(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0047941
(43) 공개일자 2019년05월09일

(51) 국제특허분류(Int. Cl.)
G06F 16/00 (2019.01)
(52) CPC특허분류
G06F 16/951 (2019.01)
G06F 16/3331 (2019.01)
(21) 출원번호 10-2017-0142349
(22) 출원일자 2017년10월30일
심사청구일자 2017년10월30일

(71) 출원인
한림대학교 산학협력단
강원도 춘천시 한림대학길 1, 한림대학교(옥천동)
(72) 발명자
김의직
강원도 춘천시 지석로 29, 514동 1103호 (석사동, 휴먼타운아파트)
권정혁
강원도 춘천시 스무숲1길 32-9, 202호 (석사동)
(74) 대리인
유민규

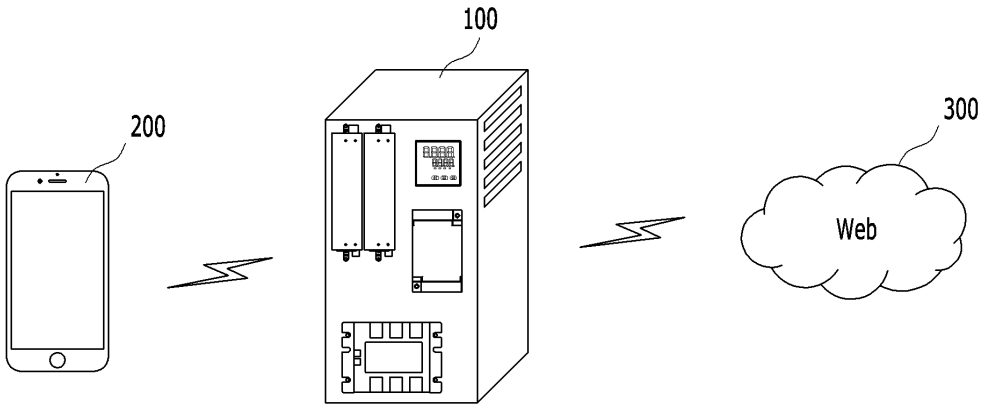
전체 청구항 수 : 총 22 항

(54) 발명의 명칭 **텍스트 데이터 수집 및 분석의 통합을 위한 방법 및 장치**

(57) 요 약

텍스트 데이터 수집 및 분석 방법은, 키워드 및 기간 정보에 기초하여 웹(web)으로부터 상기 키워드를 포함하는 기사의 정보를 획득하는 단계, 상기 기사의 정보에 기초하여 상기 기사가 포함된 웹 페이지를 크롤링(crawling)하는 단계, 상기 크롤링한 웹 페이지에 포함된 기사로부터 데이터 분석에 사용될 기반 데이터 셋트를 형성하는 단계, 상기 기반 데이터 셋트에 기초하여 텍스트 데이터 분석을 수행하는 단계 및 상기 텍스트 데이터 분석의 결과를 데이터베이스에 저장하는 단계를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/9566 (2019.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 NRF-2017R1D1A1B03031055
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 이공학개인기초연구지원사업(기본연구)
연구과제명 산업환경 안전을 위한 사물인터넷 빅데이터 통합 프레임워크 개발
기 여 율 1/1
주관기관 한림대학교 산학협력단
연구기간 2017.06.01 ~ 2018.02.28

명 세 서

청구범위

청구항 1

텍스트 데이터 수집 및 분석 방법에 있어서,

키워드 및 기간 정보에 기초하여 웹(web)으로부터 상기 키워드를 포함하는 기사의 정보를 획득하는 단계;

상기 기사의 정보에 기초하여 상기 기사가 포함된 웹 페이지를 크롤링(crawling)하는 단계;

상기 크롤링한 웹 페이지에 포함된 기사로부터 데이터 분석에 사용될 기반 데이터 셋트를 형성하는 단계;

상기 기반 데이터 셋트에 기초하여 텍스트 데이터 분석을 수행하는 단계; 및

상기 텍스트 데이터 분석의 결과를 데이터베이스에 저장하는 단계,

를 포함하는 텍스트 데이터 수집 및 분석 방법.

청구항 2

제 1 항에 있어서,

상기 키워드 및 상기 기간 정보에 관한 입력을 수신하는 단계,

를 더 포함하고,

상기 기간 정보는 기사의 게재 기간인 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 3

제 1 항에 있어서,

상기 기사의 정보를 획득하는 단계 및 상기 웹 페이지를 크롤링하는 단계는,

상기 키워드 및 상기 기간 정보를 포함하는 URL을 생성하여 상기 웹으로 전송하고, 웹으로부터 검색결과가 포함된 HTML 파일을 수신하는 것을 포함하는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 4

제 1 항에 있어서,

상기 크롤링한 웹 페이지를 미리 설정된 사전 정의 단어에 기초하여 전처리하는 단계,

를 더 포함하고,

상기 기반 데이터 셋트를 형성하는 단계는 상기 전처리된 웹 페이지의 기사로부터 상기 기반 데이터 셋트를 형성하는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 5

제 4 항에 있어서,

상기 크롤링한 웹 페이지를 전처리하는 단계는,

상기 미리 설정된 사전 정의 단어를 포함하는 기사와 상기 미리 설정된 사전 정의 단어를 포함하지 않는 기사를 분류하는 단계; 및

상기 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터 중에서 미리 설정된 텍스트 요소를 제거하는 단계,

를 포함하는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 6

제 5 항에 있어서,

상기 기반 데이터 셋트를 형성하는 단계는 상기 미리 설정된 텍스트 요소가 제거된 기사의 텍스트 데이터에 대하여 수행되는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 7

제 6 항에 있어서,

상기 기반 데이터 셋트는 상기 미리 설정된 텍스트 요소가 제거된 기사에 포함된 체언 집합, 미리 설정된 빈도 순위 이내의 빈발 단어 집합 및 각 기사의 빈발 단어 포함 여부에 관한 정보인 매트릭스 데이터 셋트를 포함하는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 8

제 1 항에 있어서,

상기 텍스트 데이터 분석을 수행하는 단계는 기사에서 사용된 단어들의 빈도수 분석 및 기사 내 단어들 간의 연관규칙 분석을 포함하는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 9

제 8 항에 있어서,

상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과를 서로 다른 그래픽으로 출력하는 단계,

를 더 포함하는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 10

제 8 항에 있어서,

상기 텍스트 데이터 분석의 결과를 저장하는 단계는, 상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과의 식별자와 연계하여 상기 텍스트 데이터 분석의 결과를 저장하는 것인, 텍스트 데이터 수집 및 분석 방법.

청구항 11

텍스트 데이터 수집 및 분석 장치에 있어서,

키워드 및 기간 정보에 기초하여 웹(web)으로부터 상기 키워드를 포함하는 기사의 정보를 획득하고, 상기 기사의 정보에 기초하여 상기 기사가 포함된 웹 페이지를 크롤링(crawling)하는 웹 크롤러부;

상기 크롤링한 웹 페이지에 포함된 기사로부터 데이터 분석에 사용될 기반 데이터 셋트를 형성하며, 상기 기반 데이터 셋트에 기초하여 텍스트 데이터 분석을 수행하는 데이터 분석부; 및

상기 텍스트 데이터 분석의 결과를 저장하는 데이터베이스,

를 포함하는 텍스트 데이터 수집 및 분석 장치.

청구항 12

제 11 항에 있어서,

상기 키워드 및 상기 기간 정보에 관한 입력을 수신하는 유저 인터페이스부,

를 더 포함하고,

상기 기간 정보는 기사의 게재 기간인 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 13

제 11 항에 있어서,

상기 웹 크롤러부는,

상기 키워드 및 상기 기간 정보를 포함하는 URL을 생성하여 상기 웹으로 전송하고, 웹으로부터 검색결과가 포함된 HTML 파일을 수신하여 상기 기사의 정보를 획득하고 상기 웹 페이지를 크롤링하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 14

제 11 항에 있어서,

상기 데이터 분석부는,

미리 설정된 사전 정의 단어에 기초하여 상기 크롤링한 웹 페이지를 전처리하는 전처리부; 및

상기 전처리된 웹 페이지의 기사로부터 상기 기반 데이터 셋트를 형성하는 데이터 형성부,

를 포함하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 15

제 14 항에 있어서,

상기 전처리부는, 상기 미리 설정된 사전 정의 단어를 포함하는 기사와 상기 미리 설정된 사전 정의 단어를 포함하지 않는 기사를 분류하고, 상기 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터 중에서 미리 설정된 텍스트 요소를 제거하여, 상기 크롤링한 웹 페이지를 전처리하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 16

제 15 항에 있어서,

상기 데이터 형성부는 상기 미리 설정된 텍스트 요소가 제거된 기사의 텍스트 데이터를 이용하여 상기 기반 데이터 셋트를 형성하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 17

제 16 항에 있어서,

상기 기반 데이터 셋트는 상기 미리 설정된 텍스트 요소가 제거된 기사에 포함된 체언 집합, 미리 설정된 빈도 순위 이내의 빈발 단어 집합 및 각 기사의 빈발 단어 포함 여부에 관한 정보인 매트릭스 데이터 셋트를 포함하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 18

제 11 항에 있어서,

상기 데이터 분석부는 상기 기반 데이터 셋트에 기초하여 기사에서 사용된 단어들의 빈도수 분석 및 기사 내 단어들 간의 연관규칙 분석을 수행하는 분석부를 포함하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 19

제 18 항에 있어서,

상기 분석부는 상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과를 상기 텍스트 데이터 수집 및 분석 장치와 연계된 사용자 단말을 통해 서로 다른 그래픽으로 출력하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 20

제 18 항에 있어서,

상기 데이터베이스는 상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과의 식별자와 연계하여 상기 텍스트 데이터 분석의 결과를 저장하는 것인, 텍스트 데이터 수집 및 분석 장치.

청구항 21

제 11 항 내지 제 20 항 중 어느 한 항의 텍스트 데이터 수집 및 분석 장치; 및

상기 텍스트 데이터 수집 및 분석 장치에 키워드 및 기간 정보에 관한 입력을 제공하고, 상기 텍스트 데이터 수집 및 분석 장치로부터 텍스트 데이터 분석의 결과를 수신하여 출력하는 사용자 단말,

을 포함하는 텍스트 데이터 수집 및 분석 시스템.

청구항 22

제 1 항 내지 제 10 항 중 어느 한 항의 방법을 컴퓨터에서 실행하기 위한 프로그램을 기록한 컴퓨터에서 판독 가능한 기록매체.

발명의 설명

기 술 분 야

[0001] 본원은 텍스트 데이터 수집 및 분석의 통합을 위한 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근 민원 분석, 뉴스 큐레이션, 선호도 조사, 추천 등의 지능형 서비스에 대한 수요가 급증하고 있다. 이러한 서비스들은 대부분 필요한 정보를 사용자에게 제공하기 위해 웹 게재기사 또는 게시글에 포함된 텍스트 데이터를 수집하고, 이를 분석한다.

[0003] 예시적으로, 휴대전화, 태블릿 PC, Wi-Fi 등 생활 무선기기 및 무선통신기술이 대중화 되면서, 이들로부터 방출되는 전자파의 인체 유해성에 대한 국민들의 우려가 급증하고 있다. 전 세계적으로 전자파 관련 민원 및 소송 사례가 증가하고 있는 추세이며, 이에 따른 민원의 원인 분석 및 효율적인 대응책 마련이 시급한 실정이다. 그러나, 기존의 전자파 관련 민원 대응방안은 지역적으로 발생하는 민원에 대한 대응만을 고려하거나 전자파에 대한 일반론적인 인체유해성 만을 고려하기 때문에, 현재 사회적으로 이슈가 되고 있는 전자파 관련 민원에 대한 신속한 대응이 어려우며, 적절한 대응방안이 되지 못할 수 있다. 이러한 문제점을 해결하기 위해서는 현재 이슈가 되고 있는 전자파 관련 민원에 대한 내용분석이 반영된 대응책이 필수적으로 마련되어야 한다. 하지만, 기존 민원 분석은 전문가에 의존적이기 때문에 긴 시간이 소요될 수 있으며, 전문가의 숙련도에 따라 정확한 분석결과를 도출해 내지 못 할 수 있다. 때문에, 고조된 국민들의 불안감을 해소하는데 오랜 기간과 많은 비용 및 인력이 소모될 뿐 아니라, 부정확한 분석결과로 인해 민원 대응에 실패할 수도 있다.

[0004] 이에, 원활한 지능형 서비스를 제공하기 위해서는 웹에 게재된 다양한 종류의 기사를 실시간으로 분석하고, 텍스트 데이터를 수집하는 웹 크롤링 기술과 수집된 데이터를 분석하는 빅데이터 분석 기술이 필수적으로 요구된다.

[0005] 본원의 배경이 되는 기술은 한국공개특허공보 제2010-0094263(공개일: 2010.08.26)호에 개시되어 있다.

발명의 내용

해결하려는 과제

[0006] 본원은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 사용자로부터 키워드를 입력받고, 웹 크롤링을 이용하여 입력 키워드 관련 인터넷 개제 기사를 수집하고, 입력 키워드 관련 텍스트 데이터 분석을 수행하며, 수집된 기사 및 분석결과를 저장 및 출력할 수 있는 데이터베이스를 포함하는 텍스트 데이터 수집 및 분석 방법을 제공하고자 한다.

[0007] 또한, 본원은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 사용자로부터 입력받은 키워드에 기반하여 키워드 텍스트 데이터 수집 및 분석 결과를 다양한 형태로 시각화하여 사용자에게 제공할 수 있는 텍스트 데이터 수집 및 분석 방법을 제공하고자 한다.

[0008] 또한, 본원은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 텍스트 데이터 수집 기술과 데이터 분석

기술을 모두 고려하는 통합 프레임워크를 제공하고자 한다.

[0009] 다만, 본원의 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들도 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0010] 상기한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본원의 일 실시예에 따르면, 텍스트 데이터 수집 및 분석 방법은, 키워드 및 기간 정보에 기초하여 웹(web)으로부터 상기 키워드를 포함하는 기사의 정보를 획득하는 단계, 상기 기사의 정보에 기초하여 상기 기사가 포함된 웹 페이지를 크롤링(crawling)하는 단계, 상기 크롤링한 웹 페이지에 포함된 기사로부터 데이터 분석에 사용될 기반 데이터 셋트를 형성하는 단계, 상기 기반 데이터 셋트에 기초하여 텍스트 데이터 분석을 수행하는 단계 및 상기 텍스트 데이터 분석의 결과를 데이터베이스에 저장하는 단계를 포함할 수 있다.

[0011] 본원의 일 실시예에 따르면, 텍스트 데이터 수집 및 분석 방법은, 상기 키워드 및 상기 기간 정보에 관한 입력을 수신하는 단계를 더 포함하고, 상기 기간 정보는 기사의 게재 기간일 수 있다.

[0012] 본원의 일 실시예에 따르면, 상기 기사의 정보를 획득하는 단계 및 상기 웹 페이지를 크롤링하는 단계는, 상기 키워드 및 상기 기간 정보를 포함하는 URL을 생성하여 상기 웹으로 전송하고, 웹으로부터 검색결과가 포함된 HTML 파일을 수신하는 것을 포함할 수 있다.

[0013] 본원의 일 실시예에 따르면, 텍스트 데이터 수집 및 분석 방법은, 상기 크롤링한 웹 페이지를 미리 설정된 사전 정의 단어에 기초하여 전처리하는 단계를 더 포함하고, 상기 기반 데이터 셋트를 형성하는 단계는 상기 전처리된 웹 페이지의 기사로부터 상기 기반 데이터 셋트를 형성할 수 있다.

[0014] 본원의 일 실시예에 따르면, 상기 크롤링한 웹 페이지를 전처리하는 단계는상기 미리 설정된 사전 정의 단어를 포함하는 기사와 상기 미리 설정된 사전 정의 단어를 포함하지 않는 기사를 분류하는 단계 및 상기 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터 중에서 미리 설정된 텍스트 요소를 제거하는 단계를 포함할 수 있다.

[0015] 본원의 일 실시예에 따르면, 상기 기반 데이터 셋트를 형성하는 단계는 상기 미리 설정된 텍스트 요소가 제거된 기사의 텍스트 데이터에 대하여 수행될 수 있다.

[0016] 본원의 일 실시예에 따르면, 상기 기반 데이터 셋트는 상기 미리 설정된 텍스트 요소가 제거된 기사에 포함된 체언 집합, 미리 설정된 빈도 순위 이내의 빈발 단어 집합 및 각 기사의 빈발 단어 포함 여부에 관한 정보인 매트릭스 데이터 셋트를 포함할 수 있다.

[0017] 본원의 일 실시예에 따르면, 상기 텍스트 데이터 분석을 수행하는 단계는 기사에서 사용된 단어들의 빈도수 분석 및 기사 내 단어들 간의 연관규칙 분석을 포함할 수 있다.

[0018] 본원의 일 실시예에 따르면, 텍스트 데이터 분석을 수행하는 단계는, 상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과를 서로 다른 그래픽으로 출력하는 단계를 더 포함할 수 있다.

[0019] 본원의 일 실시예에 따르면, 상기 텍스트 데이터 분석의 결과를 저장하는 단계는, 상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과의 식별자와 연계하여 상기 텍스트 데이터 분석의 결과를 저장할 수 있다.

[0020] 본원의 일 실시예에 따르면, 텍스트 데이터 수집 및 분석 장치는, 키워드 및 기간 정보에 기초하여 웹(web)으로부터 상기 키워드를 포함하는 기사의 정보를 획득하고, 상기 기사의 정보에 기초하여 상기 기사가 포함된 웹 페이지를 크롤링(crawling)하는 웹 크롤러부, 상기 크롤링한 웹 페이지에 포함된 기사로부터 데이터 분석에 사용될 기반 데이터 셋트를 형성하며, 상기 기반 데이터 셋트에 기초하여 텍스트 데이터 분석을 수행하는 데이터 분석부 및 상기 텍스트 데이터 분석의 결과를 저장하는 데이터베이스를 포함할 수 있다.

[0021] 본원의 일 실시예에 따르면, 텍스트 데이터 수집 및 분석 장치는 상기 키워드 및 상기 기간 정보에 관한 입력을 수신하는 유저 인터페이스부를 더 포함하고, 상기 기간 정보는 기사의 게재 기간일 수 있다.

[0022] 본원의 일 실시예에 따르면, 상기 웹 크롤러부는, 상기 키워드 및 상기 기간 정보를 포함하는 URL을 생성하여 상기 웹으로 전송하고, 웹으로부터 검색결과가 포함된 HTML 파일을 수신하여 상기 기사의 정보를 획득하고 상기 웹 페이지를 크롤링하는 것일 수 있다.

[0023] 본원의 일 실시예에 따르면, 상기 데이터 분석부는, 미리 설정된 사전 정의 단어에 기초하여 상기 크롤링한 웹

페이지를 전처리하는 전처리부 및 상기 전처리된 웹 페이지의 기사로부터 상기 기반 데이터 셋트를 형성하는 데이터 형성부를 포함할 수 있다.

[0024] 본원의 일 실시예에 따르면, 상기 전처리부는, 상기 미리 설정된 사전 정의 단어를 포함하는 기사와 상기 미리 설정된 사전 정의 단어를 포함하지 않는 기사를 분류하고, 상기 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터 중에서 미리 설정된 텍스트 요소를 제거하여, 상기 크롤링한 웹 페이지를 전처리하는 것일 수 있다.

[0025] 본원의 일 실시예에 따르면, 상기 데이터 형성부는 상기 미리 설정된 텍스트 요소가 제거된 기사의 텍스트 데이터를 이용하여 상기 기반 데이터 셋트를 형성할 수 있다.

[0026] 본원의 일 실시예에 따르면, 상기 기반 데이터 셋트는 상기 미리 설정된 텍스트 요소가 제거된 기사에 포함된 체언 집합, 미리 설정된 빈도 순위 이내의 빈발 단어 집합 및 각 기사의 빈발 단어 포함 여부에 관한 정보인 매트릭스 데이터 셋트를 포함할 수 있다.

[0027] 본원의 일 실시예에 따르면, 상기 데이터 분석부는 상기 기반 데이터 셋트에 기초하여 기사에서 사용된 단어들의 빈도수 분석 및 기사 내 단어들 간의 연관규칙 분석을 수행하는 분석부를 포함할 수 있다.

[0028] 본원의 일 실시예에 따르면, 상기 분석부는 상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과를 상기 텍스트 데이터 수집 및 분석 장치와 연계된 사용자 단말을 통해 서로 다른 그래픽으로 출력할 수 있다.

[0029] 본원의 일 실시예에 따르면, 상기 데이터베이스는 상기 빈도수 분석의 결과 및 상기 연관규칙 분석의 결과의 식별자와 연계하여 상기 텍스트 데이터 분석의 결과를 저장할 수 있다.

[0030] 본원의 일 실시예에 따르면, 텍스트 데이터 수집 및 분석 시스템은, 텍스트 데이터 수집 및 분석 장치 및 상기 텍스트 데이터 수집 및 분석 장치에 키워드 및 기간 정보에 관한 입력을 제공하고, 상기 텍스트 데이터 수집 및 분석 장치로부터 텍스트 데이터 분석의 결과를 수신하여 출력하는 사용자 단말을 포함할 수 있다.

[0031] 상술한 과제 해결 수단은 단지 예시적인 것으로서, 본원을 제한하려는 의도로 해석되지 않아야 한다. 상술한 예시적인 실시예 외에도, 도면 및 발명의 상세한 설명에 추가적인 실시예가 존재할 수 있다.

발명의 효과

[0032] 전술한 본원의 과제 해결 수단에 의하면, 사용자로부터 키워드를 입력받고, 웹 크롤링을 이용하여 입력 키워드 관련 인터넷 개제 기사를 수집하고, 입력 키워드 관련 텍스트 데이터 분석을 수행하며, 수집된 기사 및 분석결과를 저장할 수 있는 데이터베이스를 포함하는 텍스트 데이터 수집 및 분석 방법을 제공할 수 있다.

[0033] 또한, 전술한 본원의 과제 해결 수단에 의하면, 사용자로부터 입력받은 키워드에 기반하여 키워드 텍스트 데이터 수집 및 분석 결과를 다양한 형태로 시각화하여 사용자에게 제공할 수 있는 텍스트 데이터 수집 및 분석 방법을 제공할 수 있다.

[0034] 또한, 전술한 본원의 과제 해결 수단에 의하면, 웹 게제 한글 텍스트 데이터를 수집할 수 있고, 데이터 분석 자동화를 통한 분석 시간 절감 및 텍스트 데이터 수집 및 분석을 통한 통합 프레임워크를 획득할 수 있다.

도면의 간단한 설명

[0035] 도 1은 본원의 일 실시예에 따른 텍스트 데이터 수집 및 분석 시스템의 개략적인 구성을 나타낸 도면이다.

도 2는 본원의 일 실시예에 따른 텍스트 데이터 수집 및 분석 장치의 개략적인 구성을 나타낸 블록도이다.

도3은 본원의 일 실시예에 따른 텍스트 데이터 수집 및 분석 방법을 개략적으로 나타낸 흐름도이다.

도4는 본원의 일 실시예에 따른 기사의 정보를 획득하는 단계를 개략적으로 나타낸 흐름도이다.

도5는 본원의 일 실시예에 따른 기사의 텍스트 데이터를 수집하는 단계를 개략적으로 나타낸 흐름도이다.

도6는 본원의 일 실시예에 따른 수집한 텍스트 데이터 셋트를 예시적으로 나타낸 도면이다.

도7은 본원의 일 실시예에 따른 수집한 텍스트 데이터 전처리 단계를 개략적으로 나타낸 흐름도이다.

도8은 본원의 일 실시예에 따른 텍스트 데이터 분류 과정을 예시적으로 나타낸 도면이다.

도9은 본원의 일 실시예에 따른 미리 설정된 제거 텍스트 요소를 제거하는 과정을 예시적으로 나타낸 도면이다.

도10은 본원의 일 실시예에 따른 기반 데이터 셋트를 형성하는 단계를 개략적으로 나타낸 흐름도이다.

도 11은 본원의 일 실시예에 따른 매트릭스 데이터 셋트를 개략적으로 나타낸 도면이다.

도 12는 본원의 일 실시예에 따른 텍스트 데이터 분석을 수행하는 단계를 개략적으로 나타낸 흐름도이다.

도 13은 본원의 일 실시예에 따른 결정된 최소 빈도수 이상의 단어를 빈도수에 따라 출력 위치, 출력 크기 및 출력 색을 결정하여 워드클라우드 형태로 출력한 예를 나타낸 도면이다.

도14는 본원의 일 실시예에 따른 제 1단어 집합과 상기 제 2 단어 집합 간의 연관성을 매트릭스 형태로 출력하는 그룹 매트릭스 출력의 예를 나타낸 도면이다.

도15는 본원의 일 실시예에 따른 제 1단어 집합에 속한 단어와 제 2단어 집합에 속한 단어 간의 연관성을 네트워크 그래프 형태로 출력하는 그래프 출력의 예를 나타낸 도면이다.

도16은 본원의 일 실시예에 따른 결정된 연관규칙의 분포를 그래프로 출력하는 분포도 출력의 예를 나타낸 도면이다.

도17은 본원의 일 실시예에 따른 추출 빈도 및 추출 확률을 포함하는 그래프의 예를 나타낸 도면이다.

도18은 본원의 일 실시예에 따른 requestFlag 값에 따른 요청 목적을 개략적으로 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0036] 아래에서는 첨부한 도면을 참조하여 본원이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본원의 실시예를 상세히 설명한다. 그러나 본원은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본원을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0037] 아래에서는 첨부한 도면을 참조하여 본원이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본원의 실시예를 상세히 설명한다. 그러나 본원은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본원을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0038] 본원 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.

[0039] 본원 명세서 전체에서, 어떤 부재가 다른 부재 "상에", "상부에", "상단에", "하에", "하부에", "하단에" 위치하고 있다고 할 때, 이는 어떤 부재가 다른 부재에 접해 있는 경우뿐 아니라 두 부재 사이에 또 다른 부재가 존재하는 경우도 포함한다.

[0040] 본원 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함" 한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성 요소를 더 포함할 수 있는 것을 의미한다.

[0041] 본원은 텍스트 데이터 수집 및 분석 방법에 관한 것으로서, 사용자의 입력 키워드로부터 웹크롤링을 통해 웹 게재 기사 또는 게시글에 포함된 텍스트 데이터를 수집하고, 분석을 위해 수집된 데이터를 데이터베이스에 저장하며, 미리 정의된 사전 정의 단어에 따라 수집된 텍스트를 분류하고, 이에 기초하여 키워드와 추출 단어 간의 연관규칙 분석을 수행하여, 이를 시각화 기능을 사용하여 사용자에게 분석 결과를 제공할 수 있다.

[0042] 도1은 본원의 일 실시예에 따른 텍스트 데이터 수집 및 분석 시스템의 개략적인 구성을 나타낸 도면이고, 도 2는 본원의 일 실시예에 따른 텍스트 데이터 수집 및 분석 장치의 개략적인 구성을 나타낸 도면이다. 도 1 내지 도 2를 참조하면, 텍스트 데이터 수집 및 분석 시스템은, 텍스트 데이터 수집 및 분석 장치(100), 사용자 단말(200) 및 웹 (300) 포함할 수 있다. 텍스트 데이터 수집 및 분석 시스템은, 키워드 및 기간 정보에 기초하여 웹(300)으로부터 인터넷 게재 기사를 수집하고 수집된 기사에 기초하여 관련 키워드 텍스트 데이터 수집 및 분석 결과를 사용자 단말(200)에 제공하는 것이다. 사용자 단말(200)은 텍스트 데이터 수집 및 분석 장치(100)에 키워드 및 기간 정보에 관한 입력을 제공하고, 텍스트 데이터 수집 및 분석 장치(100)로부터 텍스트 데이터 분석의 결과를 수신하여 출력할 수 있다.

[0043] 사용자 단말(200)은 스마트폰(Smartphone), 스마트패드(SmartPad), 스마트 TV, 태블릿 PC등과 PCS(Personal Communication System), GSM(Global System for Mobile communication), PDC(Personal Digital Cellular),

PHS(Personal Handyphone System), PDA(Personal Digital Assistant), IMT(International Mobile Telecommunication)-2000 단말, CDMA(Code Division Multiple Access)-2000 단말, W-CDMA(W-Code Division Multiple Access) 단말, Wibro(Wireless Broadband Internet) 단말 같은 데이터 처리 및 출력이 가능한 모든 종류의 무선 통신 장치 중 어느 하나일 수 있다.

[0044] 텍스트 데이터 수집 및 분석 장치(100), 사용자 단말(200) 및 웹 (300)은 네트워크로 연결될 수 있다. 네트워크는 단말 및 서버와 같은 각각의 노드 상호 간에 정보 교환이 가능한 유, 무선의 연결 구조를 의미하는 것으로, 이러한 네트워크의 일 예에는 3GPP(3rd Generation Partnership Project) 네트워크, LTE(Long Term Evolution) 네트워크, 5G 네트워크, WIMAX(World Interoperability for Microwave Access) 네트워크, 인터넷(Internet), LAN(Local Area Network), Wireless LAN(Wireless Local Area Network), WAN(Wide Area Network), PAN(Personal Area Network), 블루투스(Bluetooth) 네트워크, 위성 방송 네트워크, 아날로그 방송 네트워크, DMB(Digital Multimedia Broadcasting) 네트워크 등이 포함되나 이에 한정되지는 않는다.

[0045] 도 1을 참조하면, 사용자 단말(200)에서 입력된 정보는 텍스트 데이터 수집 및 분석 장치(100)로 전송되어, 텍스트 데이터 수집 및 분석 장치(100)에 의해 입력 키워드 관련 텍스트 데이터가 수집 및 분석될 수 있다. 텍스트 데이터 수집 및 분석 장치(100)는 네트워크를 통해 사용자 단말(200)로 시각화한 텍스트 데이터 수집 및 분석 결과를 제공할 수 있다. 텍스트 데이터 수집 및 분석 장치(100)는 사용자의 입력 키워드 정보에 기초하여 웹(300)으로부터 입력된 키워드와 관련된 인터넷 게재 기사를 수집할 수 있다.

[0046] 텍스트 데이터 수집 및 분석 장치(100)는 키워드 및 기간 정보에 기초하여 웹(web)으로부터 키워드를 포함하는 기사 정보를 획득하고, 기사의 정보에 기초하여 기사가 포함된 웹 페이지를 크롤링(crawling)하고, 크롤링한 웹 페이지에 포함된 기사로부터 텍스트 데이터 분석을 수행하고, 분석 결과를 저장할 수 있다.

[0047] 도2를 참조하면, 텍스트 데이터 수집 및 분석 장치(100)는 크게 유저 인터페이스부(110), 웹 크롤러부(120), 데이터 분석부(130) 및 데이터베이스(140)를 포함할 수 있다. 웹 크롤러부(120)는 파싱부(121), 추출부(122), 언어 지원부(123) 및 파일 생성부(124)를 포함할 수 있다. 데이터 분석부(130)는 전처리부(131), 데이터 형성부(132) 및 분석부(133)를 포함할 수 있다. 또한, 데이터 베이스(140)는 크롤링 데이터베이스(141), 분석 데이터베이스(142) 및 사전 정의 단어 데이터베이스(143)를 포함할 수 있다.

[0048] 본원의 일 실시예에 따르면, 유저 인터페이스부(110)는 GUI를 사용자에게 제공함으로써 웹 크롤링 및 텍스트 데이터 분석을 위한 세부적인 값들을 설정하도록 도와주고, 도출된 연관규칙 분석 결과 및 워드 클라우드 (Word Cloud) 결과를 출력하여 사용자에게 제공할 수 있다. 도시하지 않았지만, 유저 인터페이스부(110)는 1) User Input Panel, 2) Result View Panel, 3) Pre-define Knowledge Panel 세개의 기능블록을 포함할 수 있다. User Input Panel은 키워드, 기사 게재기간, 페이지 및 크롤링 범위 설정 등 데이터 수집을 위한 설정과, 출력조건, Word Cloud 최소 빈도수 설정 등 데이터 분석을 위한 다양한 옵션값 설정을 위해 사용될 수 있다. Result View Panel은 데이터 분석 후 시각화된 결과를 제공하기 위해 사용되며, Pre-define Knowledge Panel은 Pre-define Knowledge DB에 저장된 단어들을 추가/삭제하기 위해 사용될 수 있다.

[0049] 웹 크롤러부(120)는 데이터 분석에 필요한 데이터 셋트를 추출하기 위해 사용자로부터 입력받은 특정 키워드 및 기사 게재기간을 조건으로 수집 가능한 기사를 검색하고, 검색된 모든 기사의 텍스트 데이터를 수집한다. 웹 크롤러부(120)는 사용자가 입력한 키워드 및 옵션 값들을 기반으로 URL을 생성하고, 생성한 URL을 사용하여 웹에 게재된 기사를 검색한다. 이 후 검색된 기사의 목록 및 기사 본문을 포함하는 HTML 파일을 웹(300)으로부터 가져온다. 또한, 웹 크롤러부(120)는 한글 지원을 위한 인코딩을 수행하고, HTML 파일 내용 중 기사의 제목 및 내용에 해당하는 텍스트 데이터를 추출하여, 추출된 텍스트 데이터들을 txt 형식으로 생성 및 저장할 수 있다.

[0050] 데이터 분석부(130)는 웹 크롤러부(120)의 수행 결과로 수집된 데이터 셋트를 기반으로 데이터 전처리를 수행하고, 연관규칙 분석을 위한 매트릭스 데이터를 형성한다. 이후, 연관규칙 분석을 수행하고, 분석 결과를 다양한 형태로 시각화할 수 있다. 데이터 분석부(130)는 수집된 기사들의 데이터를 미리 설정된 단어와 비교하여 분석할 기사를 분류한다. 또한, 분류된 기사의 본문으로부터 체언 (명사)을 추출하고, 추출된 단어 중 빈도수가 높은 단어들을 기반으로 연관규칙 분석을 위한 데이터 셋트를 형성할 수 있다. 이후, 데이터 분석부(130)는 Apriori 알고리즘을 활용한 연관규칙 분석을 수행하고, 분석결과를 Word Cloud, Grouped Matrix, Graph, Scatter Plot, Extraction Probability 형태로 시각화할 수 있다.

[0051] 데이터베이스(140)는 웹 크롤러부(120)에 의해 추출된 데이터 셋트를 저장하고, 데이터 분석부(130)를 통해 얻은 분석 결과 및 시각화된 이미지를 저장할 수 있다. 특히, 데이터 전처리 과정에서 수행되는 기사 분류 작업을

위해 사용자가 정의한 미리 설정된 단어들을 저장하며, 전처리 결과 및 연관분석을 위한 데이터 셋트를 저장한다. 데이터베이스(140)는 1) 크롤링 데이터베이스, 2) 분석 데이터베이스, 3) 사전 정의 단어 데이터베이스를 포함한다. 크롤링 데이터베이스는 웹 크롤러부(120)를 통해 수집된 데이터를 저장 및 관리를 위해 사용되며, 분석 데이터베이스는 데이터 분석 결과를 저장 및 관리하기 위해, 그리고 사전 정의 단어 데이터베이스는 사용자 정의 기반의 사전 정의 단어를 저장 및 관리하기 위해 사용될 수 있다.

[0052] 텍스트 데이터 수집 및 분석 장치(100)의 각 부에 대한 설명은 아래에서 자세히 설명하도록 한다.

[0053] 도3은 본원의 일 실시예에 따른 텍스트 데이터 수집 및 분석 방법을 개략적으로 나타낸 흐름도이다. 또한, 도 3에 도시된 각 단계의 설명은 도 4내지 도 12를 통해 구체적으로 설명하도록 한다.

[0054] 도3을 참조하면, 단계 S310에서, 유저 인터페이스부(110)는 웹 크롤러부(120)로 키워드 및 기간 정보를 전송하고, 웹 크롤러부(120)는 키워드 및 기간 정보를 수신할 수 있다. 예를 들어, 기간 정보는 기사의 게재 기간일 수 있다.

[0055] 단계 S320에서, 웹 크롤러부(120)는 키워드 및 기간 정보에 기초하여 웹(web)으로부터 키워드를 포함하는 기사의 정보를 획득할 수 있다. 또한, 단계 S330에서, 웹 크롤러부(120)는 기사의 정보에 기초하여 기사가 포함된 웹 페이지를 크롤링(crawling)할 수 있다. 단계 S320 및 단계330에서, 웹 크롤러부(120)는 키워드 및 기간 정보를 포함하는 URL을 생성하여 웹으로 전송하고, 웹으로부터 검색결과가 포함된 HTML 파일을 수신할 수 있다.

[0056] 단계 S340에서, 웹 크롤러부(120)는 데이터 분석부(130)로 크롤링한 웹 페이지에 포함된 기사의 텍스트 데이터를 전송할 수 있다. 또 다른 예로서, 단계 S340에서, 데이터 분석부(130)는 데이터베이스로부터 웹 크롤러부(120)에 의해 수집된 기사의 텍스트 데이터를 읽어올 수 있다.

[0057] 단계 S350에서, 데이터 분석부(130)는 미리 설정된 사전 정의 단어에 기초하여 크롤링한 웹 페이지를 전처리 할 수 있다. 보다 구체적으로, 데이터 분석부(130)는 미리 설정된 사전 정의 단어에 기초하여 크롤링한 웹 페이지에 포함된 기사의 텍스트 데이터를 전처리할 수 있다. 데이터 분석부(130)는 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터와 미리 설정된 사전 정의 단어를 포함하지 않는 기사의 텍스트 데이터를 분류할 수 있다. 또한, 데이터 분석부(130)는 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터 중에서 미리 설정된 제거 텍스트 요소를 제거할 수 있다.

[0058] 단계 S360에서, 데이터 분석부(130)는 크롤링한 웹 페이지에 포함된 기사로부터 데이터 분석에 사용될 기반 데이터 셋트를 형성할 수 있다. 보다 구체적으로, 데이터 분석부(130)는 전처리된 웹 페이지의 기사의 텍스트 데이터로부터 기반 데이터 셋트를 형성할 수 있다. 데이터 분석부(130)는 미리 설정된 텍스트 제거 요소가 제거된 기사의 텍스트 데이터에 대하여 기반 데이터 셋트를 형성할 수 있다. 본원의 일 실시예에 따르면, 상기 기반 데이터 셋트는 미리 설정된 텍스트 제거 요소가 제거된 기사에 포함된 체언 집합, 미리 설정된 순위 이내의 빈발 단어 집합 및 각 기사의 빈발 단어 포함 여부에 관한 정보인 매트릭스 데이터 셋트를 포함할 수 있다.

[0059] 단계 S370에서, 데이터 분석부(130)는 기반 데이터 셋트에 기초하여 텍스트 데이터를 분석할 수 있다. 본원의 일 실시예에 따르면, 데이터 분석부(130)는 기사에서 사용된 단어들의 빈도수 분석 및 기사 내 단어들 간의 연관규칙 분석을 포함하여 텍스트 데이터 분석을 수행할 수 있다. 또한, 데이터 분석부(130)는 빈도수 분석의 결과 및 연관규칙 분석의 결과를 서로 다른 그래픽으로 출력할 수 있다.

[0060] 단계S380에서, 데이터 베이스(140)는 텍스트 데이터 분석의 결과를 저장할 수 있다. 본원의 일 실시예에 따르면, 데이터 베이스(140)는 빈도수 분석의 결과 및 연관규칙 분석의 결과의 식별자와 연계하여 상기 텍스트 데이터 분석의 결과를 저장할 수 있다. 예를 들어, 연관규칙 분석의 결과의 식별자는 Word Cloud, Grouped Matrix, Graph, Scatter Plot, Extraction Probability 등을 포함할 수 있다.

[0061] 도4는 본원의 일 실시예에 따른 기사의 정보를 획득하는 단계를 개략적으로 나타낸 흐름도이다.

[0062] 도4를 참조하면, 단계 S410에서 유저 인터페이스부(110)는 키워드 및 기간 정보에 관한 입력을 수신할 수 있다. 기간 정보는 기사의 게재 기간일 수 있다. 예를 들어, 사용자는 기사 게재 기간에 제한을 두지 않고 검색을 수행하도록 설정할 수 있고, 특정 기간 내에 게재된 기사를 검색하고자 하는 경우, 시작 일자(년,월,일) 및 마지막 일자(년,월,일)를 각각 입력하여 게재기간을 설정할 수 있다. 예를 들어, 유저 인터페이스부(110)는 캘린더 UI를 사용하여 사용자가 게재기간의 시작일자 및 마지막 일자를 설정할 수 있게 할 수 있다.

[0063] 단계 S420에서, 유저 인터페이스부(110)는 키워드 및 기간 정보와 함께 요청 정보를 포함하는 요청 플래그(RequestFlag)를 웹 크롤러부(120)로 전송할 수 있다. 예시적으로, 도 18을 참조하면, 요청 1(request 1)은 검

색된 기사의 수에 대한 요청일 수 있다. 유저 인터페이스부(110)는 요청 1(request 1)을 포함하는 요청 플래그를 전송할 수 있다. 즉, 유저 인터페이스부(110)는 키워드, 기간 정보 및 요청 1을 웹 크롤러부(120)로 전송할 수 있다.

[0064] 웹 크롤러부(120)는 유저 인터페이스부(110)로부터 전송받는 요청1의 목적인 검색된 기사의 수를 계산할 수 있다.

[0065] 단계 S430에서, 웹 크롤러부(120)의 파싱부(121)는 키워드, 기간 정보 및 웹 페이지의 정보를 포함하는 URL을 생성할 수 있다. 일예로, 파싱부(121)는 기간 정보가 특정되어 있지 않을 경우 기사의 게재 기간을 한정하지 않고, 기사 검색결과를 얻기 위한 URL을 생성할 수 있다. 다른 일예로, 파싱부(121)는 기간 정보가 시작 일자 및 마지막 일자가 설정되어 있다면, 특정 값으로 설정된 게재기간을 기반으로 기사 검색결과를 얻기 위한 URL을 생성할 수 있다.

[0066] 단계S440에서, 파싱부(121)는 생성한 URL을 웹(300)으로 전송하고, 웹(300)에 검색 결과가 포함된 HTML 파일을 웹(300)요청할 수 있다.

[0067] 단계 S450에서, 웹(300)은 키워드 및 기간 정보를 포함하는 HTML 파일을 생성하고, 단계S460에서, 웹(300)은 검색 결과가 포함된 HTML 파일을 웹 크롤러부(120)로 전송할 수 있다.

[0068] 단계S470에서, 파싱부(121)는 웹(300)으로부터 수신된 HTML 파일을 트리 형태로 재구성할 수 있다. 추출부(122)는 재구성된 HTML 파일에서 기사의 수 추출과 연계된 미리 설정된 속성 값(예를 들어, span class="result_num")을 가지는 데이터를 추출할 수 있다.

[0069] 단계S480에서, 추출부(122)는 추출된 데이터로부터 공백 및 특수문자를 제거하고 기사의 수를 집계할 수 있다. 예시적으로, 웹 크롤러부(120)는 추출이 완료된 데이터는 검색된 기사의 수를 “total: 기사의 수”의 형태로 유저 인터페이스부(100)로 전송할 수 있다. 웹 크롤러부(120)는 검색결과가 존재하지 않는면, “total: NA”의 형태로 유저 인터페이스부(120)로 전송할 수 있다.

[0070] 단계 S490에서, 웹 크롤러부(120)는 기사의 수에 기초하여 크롤링 가능한 웹 페이지의 수를 계산할 수 있다. 예를 들어, 하나의 웹 페이지에는 적어도 하나의 기사가 포함될 수 있다. 웹 크롤러부(120)는 키워드 및 기간 정보에 기초하여 키워드를 포함하는 기사의 결과 즉 검색 결과가 존재한다면 최대 크롤링 가능한 웹 페이지의 수를 숫자로 표현하여 유저 인터페이스부(120)로 전송할 수 있다. 웹 크롤러부(120)는 검색 결과가 존재하지 않는다면 텍스트 형태로 표현하여 검색 결과가 없음을 유저 인터페이스부(120)로 전송할 수 있다.

[0071] 도5는 본원의 일 실시예에 따른 기사의 텍스트 데이터를 수집하는 단계를 개략적으로 나타낸 흐름도이다.

[0072] 웹 크롤러부(120)는 유저 인터페이스부(110)에서 사용자 등에 의해 입력 받은 웹 페이지의 범위 내에서 키워드를 포함하는 기사가 포함된 웹 페이지를 크롤링 할 수 있다.

[0073] 단계S510에서, 유저 인터페이스부(110)는 사용자 등에 의해 크롤링 범위 입력을 수신할 수 있다. 크롤링 범위는 크롤링 할 기사의 웹 페이지 범위를 의미하며, 상기 단계 S490에서 산출된 크롤링 가능 웹 페이지의 수 보다 작은 범위로 설정되어야 한다. 유저 인터페이스부(110)는 최대 크롤링 가능한 웹 페이지의 수를 제공할 수 있다. 사용자는 시작(start) 웹 페이지 및 종료(end) 페이지를 입력하여 크롤링 범위를 설정할 수 있다.

[0074] 단계 S520에서, 유저 인퍼페이스부(110)는 키워드와 함께 웹 크롤링을 수행하는 시작 웹 페이지를 전송할 수 있다. 또한, 유저 인터페이스부(110)는 웹 크롤러부(120)로 웹 페이지 범위 전송 시 요청 2(request 2)을 포함하는 요청 플래그를 전송할 수 있다. 도 18을 참조하면, 요청2는 기사 크롤링 요청을 목적으로 하는 요청 플래그일 수 있다. 웹 크롤러부(120)는 유저 인터페이스부(110)의 요청 2에 따라, 기사 크롤링을 수행할 수 있다.

[0075] 단계 S530에서, 웹 크롤러부(120)는 키워드 및 기간 정보에 기초하여 키워드를 포함하는 기사의 수를 계산할 수 있다. 파싱부(121)는 기사의 수를 계산하기 위해, 특정 페이지 크롤링을 위한 URL을 생성할 수 있다. URL 생성 과정은 단계 S430에서의 키워드 및 기간 정보를 포함하는 URL을 생성하는 과정과 유사하다. 파싱부(121)는 키워드, 기간 정보 및 웹 페이지의 정보를 포함하는 URL 을 생성하여 웹(300)으로 전송하고, 웹(300)으로부터 웹 페이지에 포함된 기사 리스트가 포함된 HTML 파일을 수신하며, HTML 파일을 트리형태로 재구성할 수 있다.

[0076] 단계 S540에서, 웹 크롤러부(120)는 입력 받은 웹 페이지의 범위 내에서 키워드를 포함하는 기사를 크롤링 할 수 있다. 단계 S540에서, 파싱부(121)는 상기 HTML 파일로부터 기사 리스트의 URL을 추출하여 상기 기사 리스트의 URL을 웹(300)으로 전송하고, 웹으로부터 기사 텍스트 데이터를 포함하는 HTML 파일을 수신하고, 상기 기사

텍스트 데이터를 포함하는 HTML 파일을 트리형태로 재구성할 수 있다.

[0077] 또한, 단계 S540에서, 웹 크롤러부(120)의 추출부(122)는 트리형태로 재구성된 기사 리스트의 HTML 파일로부터 텍스트 데이터를 추출할 수 있다. 또한, 추출부(122)는 상기 추출된 텍스트 데이터로부터 기사의 제목 및 내용에 해당하는 텍스트 데이터를 추출할 수 있다. 언어 지원부(123)는 트리형태로 재구성된 기사 텍스트 데이터를 포함하는 HTML 파일에 대하여 인코딩을 수행할 수 있다. 예를 들어, 인코딩은 한글 텍스트 데이터 손실을 방지하기 위해 UTP-8(Universal Transformation Format-8) 인코딩일 수 있다.

[0078] 단계 S550에서, 웹 크롤러부(120)는 기사 리스트의 수만큼 크롤링 과정을 반복하여 수행할 수 있다. 웹 크롤러부(120)는 현재 크롤링을 수행하고 있는 기사의 수가 총 기사의 수보다 이상인 경우, 반복 과정을 종료하고, 크롤링이 완료된 기사 텍스트 데이터를 크롤링 데이터 베이스로 전송할 수 있다.

[0079] 단계 S560에서, 웹 크롤러부(120)는 기사 리스트의 URL 수만큼 반복하여 특정 페이지의 기사 텍스트 데이터 요청 과정이 완료되면 기사 텍스트 데이터를 크롤링 데이터베이스(300)로 전송할 수 있다. 유저 인터페이스부(110)는 크롤링 데이터베이스(141)에 기사 텍스트 데이터의 저장이 완료되면, 사용자가 지정한 웹 페이지 범위 중 시작 페이지와 종료 페이지를 비교할 수 있다.

[0080] 단계 S570에서 웹 크롤러부(120)는 종료 페이지의 수가 현재 크롤링을 수행하고 있는 웹 페이지의 수보다 크다면 웹 크롤러부(120)는 웹 페이지의 시작 페이지를 1로 증가시키고 웹 페이지에 포함되어 있는 기사의 텍스트 데이터의 크롤링 과정을 반복하여 수행할 수 있다. 즉, 웹 크롤러부(120)는 입력 받은 웹 페이지의 범위 내에서 웹 페이지의 크롤링이 반복하여 수행될 수 있다.

[0081] 또한, 파일 생성부(124)는 추출된 텍스트 데이터를 txt 형식의 파일로 생성할 수 있다. 크롤링 데이터 베이스(141)는 생성된 txt 형식의 파일을 저장할 수 있다.

[0082] 도6는 본원의 일 실시예에 따른 수집한 텍스트 데이터 셋트를 예시적으로 나타낸 도면이다.

[0083] 도6을 참조하면, 크롤링된 웹 페이지에 포함된 기사의 텍스트 데이터를 수집 결과로, 수집한 텍스트 데이터들은 txt 형식의 파일로 저장될 수 있다. 크롤링 과정 중 언어 지원부(123)에서 기사 텍스트 데이터를 포함하는 HTML 파일에 대하여 인코딩을 수행함으로써, 한글 데이터가 손실 되는 것을 방지 하고, 한글 형태의 기사의 제목을 얻을 수 있다.

[0084] 도7은 본원의 일 실시예에 따른 수집한 텍스트 데이터 전처리 단계를 개략적으로 나타낸 흐름도이다.

[0085] 도7을 참조하면, 전처리부(131)는 수집한 텍스트 데이터 셋트에 대하여 미리 설정된 사전 정의 단어에 기초하여 전처리를 수행할 수 있다. 데이터 분석부(130)는 요청 3및 4의 정보가 수신되면, 데이터 분석의 정확도를 높이기 위해 데이터 전처리를 수행할 수 있다. 전처리 과정은 데이터 분류 과정 및 데이터 필터링 과정으로 진행될 수 있다. 데이터 필터링 과정은 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터 중에서 미리 설정된 제거 텍스트 요소를 제거하는 과정일 수 있다.

[0086] 데이터 전처리를 위한 사전 작업으로, 사전 정의 단어 데이터베이스(143)에 사전 정의 단어를 정의할 수 있다. 사전 정의 단어는 사용자에 의해 정의될 수 있다.

[0087] 단계 S710에서, 유저 인터페이스부(110)는 사용자가 추가한 사전 정의 단어 입력을 수신할 수 있다. 사용자는 사전 정의 단어를 필요에 따라 추가하거나, 삭제할 수 있다. 유저 인터페이스부(110)는 추가 및 삭제에 대한 입력을 수신하고, 업데이트 된 사전 정의 단어에 기초하여 사전 정의 단어 데이터베이스(143)로 전송할 수 있다.

[0088] 또한, 단계 S710에서, 유저 인터페이스부(100)는 사전 정의 단어 전송 시 요청 3(request 3)을 포함하는 요청 플래그를 전송할 수 있다. 요청 3은 데이터 분류 요청을 목적으로하는 requestFlag 값일 수 있다.

[0089] 단계 S720에서, 데이터 분석부(130)의 요청으로 크롤링 데이터베이스(141)는 수집한 텍스트 데이터를 데이터 분석부(130)로 전송할 수 있다. 데이터 분석부(130)의 전처리부(131)는 텍스트 데이터를 분류하기 위해 크롤링 데이터베이스(141)에 수집된 텍스트 데이터를 전송 받을 수 있다.

[0090] 단계 S730에서, 전처리부(131)는 수집한 텍스트 데이터 중 기사 본문의 내용과 사전 정의 단어를 비교할 수 있다. 전처리부(131)는 사전 정의 단어가 포함되어 있다면, 임시 파일을 생성하여 데이터 셋트를 일차적으로 분류할 수 있다.

[0091] 단계 S740에서, 전처리부(131)는 데이터 분류 작업의 속도 향상과 텍스트 데이터 셋트의 중복 분류를 방지하기

위해 분류된 데이터는 크롤링 데이터베이스(141)에 텍스트 데이터 삭제 요청을 할 수 있다.

[0092] 즉, 전처리부(131)는 미리 설정된 사전 정의 단어를 포함하지 않는 기사의 텍스트 데이터는 크롤링 데이터베이스(141)에 유지하고, 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터를 크롤링 데이터베이스에서 삭제하고 분석 데이터베이스(142)에 저장할 수 있다.

[0093] 단계 S750에서, 크롤링 데이터베이스(141)는 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터를 삭제할 수 있다. 또한, 단계 S760에서, 분석 데이터베이스(142)는 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터를 저장한다.

[0094] 전처리부(131)는 수집한 텍스트 데이터에 대하여 미리 설정된 사전 정의 단어의 수만큼 반복하여 기사 분류를 수행할 수 있다.

[0095] 도 8은 본원의 일 실시예에 따른 텍스트 데이터 분류 과정을 예시적으로 나타낸 도면이다.

[0096] 도8(a)는 사용자가 미리 설정한 사전 정의 단어일 수 있다. 사용자는 분류하고 싶은 단어, 즉, 기사에 포함되어야 하는 단어를 사전 정의 단어로 지정하고, 수집된 텍스트 데이터로부터 사전 정의 단어의 포함 여부를 판단할 수 있다. 예를 들어, 전처리부(131)는 유저 인터페이스부(110)로부터 사용자가 미리 설정한 사전 정의 단어를 전파, 부정, 인체, 부작용, 노출이라고 정의한 입력을 수신할 수 있다. 전처리부(131)는 크롤링 데이터베이스에 수집된 텍스트 데이터 셋트 중, 사용자가 미리 설정한 사전 정의 단어를 전파, 부정, 인체, 부작용, 노출이라는 단어가 포함된 기사의 데이터를 크롤링 데이터베이스(141)에서 삭제하고, 미리 설정한 사전 정의 단어가 포함된 기사의 텍스트 데이터를 분석 데이터베이스(142)에 저장할 수 있다. 도8(b)는 텍스트 데이터 셋트로부터 미리 설정한 사전 정의 단어가 포함되지 않은 단어를 삭제하는 과정을 나타내고, 도8(c)는 전처리부(131)에서 수행된 전처리된 텍스트 데이터의 셋트의 예시일 수 있다.

[0097] 또한, 전처리부(131)는 유저 인터페이스부(110)로부터 요청 4(request 4)을 포함하는 요청 플래그를 수신할 수 있다. 요청4는 텍스트 데이터 필터링 요청을 목적으로 하는 requestFlag 값일 수 있다. 단계 S770에서, 전처리부(131)는 텍스트 데이터 필터링을 수행하기 위해 분석 데이터베이스(142)로부터 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터를 수신할 수 있다.

[0098] 단계 S780에서, 요청 4를 수신한 전처리부(131)는 분석 데이터베이스(142)에 저장된 사전 정의 단어를 포함하는 텍스트 데이터 중에서 미리 설정된 제거 텍스트 요소를 제거하는 과정을 수행할 수 있다. 예를 들어, 미리 설정된 제거 텍스트 요소는 기사를 작성한 기자의 이름, 기자의 이메일 주소, 광고 및 저작관 표기 중 적어도 어느 하나를 포함하는 것일 수 있다. 미리 설정된 제거 텍스트 요소를 제거하여, 텍스트 데이터 분석에 불필요한 요소들을 제거할 수 있다.

[0099] 구체적으로, 전처리부(131)는 미리 설정된 제거 텍스트 요소를 제거하는 과정을 수행할 때, 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터 중 마침표의 모든 위치를 검색하고, 마침표의 위치 검색 결과에 기초하여 기사의 마지막 문장을 결정하고, 결정된 마지막 문장의 다음 텍스트 영역을 미리 설정된 제거 텍스트 요소로서 제거할 수 있다.

[0100] 도9은 본원의 일 실시예에 따른 미리 설정된 제거 텍스트 요소를 제거하는 과정을 예시적으로 나타낸 도면이다.

[0101] 도9를 참조하면, 미리 설정된 사전 정의 단어를 포함하는 기사의 텍스트 데이터에는 마지막 문장 다음으로 기자 이름, 기자 이메일, 광고, 저작권 표기 등과 같은 요소들이 저장되어 있을 수 있다. 전처리부(131)는 분류된 데이터들 중 “.”의 모든 위치를 검색하고, 검색되어 나온 위치 중 끝에서 두 번째에 위치하는 곳을 기준으로 텍스트 요소를 제거할 수 있다. 예를 들어, 일반적으로 기사의 내용 중 마지막 문장의 마침표 이후에 기자의 이름, 이메일 주소, 저작권 표기 등이 기재되는데, 이메일 주소에 마침표가 포함되어 있으므로, 검색되어 나온 마침표의 위치 중 끝에서 두 번째에 위치하는 곳을 기사의 마지막 문장이 끝난 지점으로 결정하고, 마지막 문장의 다음에 위치하는 텍스트 영역을 제거할 수 있다. 전처리부(131)는 기준점으로 정한 이후의 내용을 삭제하고, 삭제가 완료된 텍스트 데이터를 분석 데이터베이스(142)에 저장할 수 있다.

[0102] 도10은 본원의 일 실시예에 따른 기반 데이터 셋트를 형성하는 단계를 개략적으로 나타낸 흐름도이다.

[0103] 도10을 참조하면, 데이터 형성 단계는 데이터 분석에 사용될 기반 데이터 셋트를 형성하기 위한 단계로, 체언 추출, 빈발단어 추출, 매트릭스 데이터 셋트 형성 과정이 순차적으로 수행될 수 있다.

[0104] 먼저, 데이터 분석부(130)의 데이터 형성부(132)는 유저 인터페이스부(110)로 요청 5(request 5)를 포함하는 요

청 플래그를 수신할 수 있다. 요청 5는 체언 추출 요청을 목적으로하는 requestFlag 값일 수 있다. 체언 추출 과정은 수집된 기사로부터 대명사, 명사, 수사와 같은 체언을 추출하는 과정일 수 있다.

[0105] 단계 S810에서, 데이터 형성부(132)는 분석 데이터 베이스(142)로부터 전처리된 기사의 텍스트 데이터를 전송받을 수 있다.

[0106] 단계S820에서, 데이터 형성부(132)는 전처리된 기사의 텍스트 데이터를 기반으로 데이터 분석에서 불필요한 조사, 형용사, 부사, 등을 제외한 체언만을 추출할 수 있다. 본원의 일 실시예에 따르면, 데이터 형성부(132)는 체언 추출 과정을 통해 추출된 체언으로 구성된 단어 셋트 중 두 글자 이상의 단어를 분류하여 분석 데이터 베이스(142)에 저장 할 수 있다.

[0107] 체언 추출 과정 후 데이터 형성부(132)는 유저 인터페이스부(110)로 요청 6(request 6)를 포함하는 요청 플래그를 수신할 수 있다. 요청6은 빈발단어 추출 요청을 목적으로하는 requestFlag 값일 수 있다. 빈발단어 추출은 추출된 체언의 빈도수를 기준으로 정렬하여 특정 순위 이상의 체언을 추출할 수 있다.

[0108] 단계 S830에서, 유저 인터페이스부(110)는 사용자의 단어 빈도수 순위에 관한 입력을 수신할 수 있다. 단어 빈도수 순위는 빈발 단어를 추출하기 위한 단어의 빈도수 순위 기준을 의미하며, 빈발단어 추출 과정에서 특정 빈도수 순위 이내인 단어를 추출하는데 사용될 수 있다. 데이터 형성부(132)는 분석 데이터베이스(142)로 체언 집합 데이터를 요청할 수 있다.

[0109] 단계 S840에서, 분석 데이터 베이스(142)는 추출된 체언 집합 데이터 셋트를 전송할 수 있다.

[0110] 단계 S850에서, 데이터 형성부(132)는 체언 집합 데이터 및 단어 빈도수 순위에 기반하여 빈발단어를 추출할 수 있다. 데이터 형성부(132)는 체언 집합 데이터 셋트 중 빈도수가 사용자가 입력한 단어 빈도수 순위 값 이내인 단어들만 추출하여 분석 데이터 베이스(142)에 저장할 수 있다.

[0111] 빈발 단어 추출 과정 후 데이터 형성부(132)는 유저 인터페이스부(110)로 요청 7(request 7)를 포함하는 요청 플래그를 수신할 수 있다. 요청7은 매트릭스 데이터 셋트 형성 요청을 목적으로하는 requestFlag 값일 수 있다. 매트릭스 데이터 셋트 형성 과정은 연관규칙 분석 및 단어 추출 확률을 위해 각 기사의 빈발단어 포함여부가 나타난 데이터 셋트를 형성하는 과정일 수 있다.

[0112] 단계 S860에서, 데이터 형성부(132)는 분석 데이터베이스(142)로부터 전처리된 기사의 텍스트 데이터 셋트 및 빈발 단어 집합을 수신할 수 있다.

[0113] 단계 S870에서 데이터 형성부(132)는 전처리된 기사의 텍스트 데이터 셋트, 빈발단어 집합 및 요청7의 요청 플래그 수신 시 포함된 키워드 값에 기반하여 연관 규칙 분석을 위해 각 기사의 빈발단어 포함여부를 나타내는 매트릭스 데이터 셋트를 형성할 수 있다. 단계 S880에서, 데이터 형성부(132)는 체언 집합, 빈발 단어 집합 및 매트릭스 데이터 셋트를 포함하는 기반 데이터 셋트를 형성한다.

[0114] 도 11은 본원의 일 실시예에 따른 매트릭스 데이터 셋트를 개략적으로 나타낸 도면이다.

[0115] 도 11을 참조하면, 매트릭스 데이터 셋트의 열 이름은 특정 빈도수 이상의 단어를 나타내고, 행 이름은 기사의 식별자로서, 1부터 기사의 수만큼의 숫자를 나타낼 수 있다. 데이터 형성부(132)는 각각의 기사와 특정 빈도수 이상의 단어를 비교하여 해당 기사가 특정 빈도수 이상의 단어를 포함하고 있으면 1로 설정하고 포함하지 않으면 0으로 설정할 수 있다. 또한, 데이터 형성부(132)는 사용자가 입력한 키워드와 추출된 단어들 간의 관계를 분석하기 위해 첫 번째 열의 이름에 사용자가 입력한 키워드를 넣고 첫 번째 열에 해당하는 행의 값을 1로 넣어준다. 사용자가 입력한 키워드는 단계S310에서 수신한 키워드 정보일 수 있다.

[0116] 도 12는 본원의 일 실시예에 따른 텍스트 데이터 분석을 수행하는 단계를 개략적으로 나타낸 흐름도이다.

[0117] 도12를 참조하면, 데이터 분석 단계는 추출된 데이터 셋트를 분석하여 사용자에게 분석 결과를 다양한 시각화 형태로 제공해주기 위한 단계일 수 있다. 데이터 분석에서는 기사에서 사용된 단어들의 빈도수 분석과 기사 내 단어들 간의 연관규칙 분석이 수행될 수 있다. 단어 빈도수 분석은 특정 단어의 사용 횟수 분석과 특정 단어의 추출확률 분석이 수행될 수 있다. 연관규칙 분석은 기사 내에 사용된 단어들의 상관관계를 분석하는 것으로, 매트릭스 데이터 셋트를 기반으로 수행될 수 있다.

[0118] 분석부(133)는 기사에서 사용된 단어들의 빈도수 분석 및 기사 내 단어들 간의 연관규칙을 분석할 수 있다. 분석부(130)는 빈발단어로 만들 수 있는 모든 단어 집합들과 각 기사가 포함하고 있는 단어집합을 비교하여, 다수의 기사에서 동시에 사용된 단어집합을 찾아 IF-THEN 형태의 규칙으로 정의하는 분석 방법을 사용하여 연관규칙

분석을 수행할 수 있다.

[0119] 예시적으로, 연관규칙은, {전자파, 중계기} -> {반대}로 정의 및 표현될 수 있다. 즉, 기사에서 '전자파, 중계기' 라는 단어가 사용되면, '반대' 라는 단어도 함께 사용된다고 해석될 수 있다. 여기서, {전자파, 중계기}는 제1단어 집합(Left-Hand Side (LHS))에 해당하며, {반대}는 제 2 단어 집합(Right-Hand Side (RHS))에 해당한다. 제1단어 집합(LHS)은 매트릭스 데이터 셋트 내 빈발단어들로 만들 수 있는 단어집합이며, 제2 단어 집합(RHS)은 매트릭스 데이터 셋트 내 빈발단어들 중 하나의 단어로 구성된 단어집합이다.

[0120] 또한, 매트릭스 데이터 셋트에 전자파, 중계기, 송전탑 세 빈발단어를 포함하고 있으면, 제 1 단어 집합(LHS)은 {전자파}, {중계기}, {송전탑}, {전자파, 중계기}, {전자파, 송전탑}, {중계기, 송전탑}, {전자파, 중계기, 송전탑}이 되며, 제 2 단어 집합(RHS)는 {전자파}, {중계기}, {송전탑}이 된다. 연관규칙에서 제1단어집합(LHS)과 제2단어 집합(RHS)은 중복 단어를 사용하지 않는다. 분석부(133)는 제1단어집합(LHS)과 제2단어 집합(RHS)에 관련된 수많은 연관 규칙들을 분석 결과로 생성할 수 있다. 제2단어 집합(RHS)을 특정단어로 사용자가 지정하면, 사용자가 원하는 단어에 대한 연관규칙 분석 결과를 쉽게 확인할 수 있다.

[0121] 연관 규칙 분석은 매트릭스 데이터 셋트 내 단어들로 만들 수 있는 모든 집합을 각 기사별로 비교하는 작업이기 때문에, 단어와 기사의 수가 많아지면 연산 횟수가 기하급수적으로 증가할 수 있다. 분석부(133)는 연관규칙 분석 시 연산 횟수를 줄이기 위해 두 가지 원칙을 기반으로 비빈발 단어집합에 대한 연산을 제거하는 Apriori알고리즘을 사용할 수 있다.

[0122] Apriori알고리즘은 1)한 항목집합이 빈발(frequent)하다면 이 항목집합의 모든 부분집합 역시 빈발항목집합이다. 2) 한 항복집합이 비빈발(infrequent)하다면 이 항목집합을 포함하는 모든 집합은 비빈발 항목집합이다. 의 두 가지 알고리즘을 포함할 수 있다.

[0123] 항목집합은 데이터 셋트(집합) 내 단어들로 만들 수 있는 집합이므로, Apriori알고리즘을 적용하면 비교적 단어 집합을 줄여 연산 횟수를 감소시킬 수 있다.

[0124] 분석부(133)는 연관규칙 분석을 통해 키워드와 추출단어 간 관계를 설정하기 위해 지지도(Support), 신뢰도(Confidence), 향상도(Lift)라는 세 가지 요소를 사용할 수 있다. 지지도, 신뢰도, 향상도는 아래와 같이 계산되며, X는 제1단어집합(LHS), Y는 제 2 단어집합(RHS)를 의미한다.

[0125] ① 지지도 (support, s(X→Y))

[0126] = X와 Y를 동시에 포함하는 기사 수 / 전체 기사 수

[0127] ② 신뢰도 (Confidence, c(X→Y))

[0128] = X와 Y를 동시에 포함하는 기사 수 / X를 포함하는 기사 수

[0129] ③ 향상도 (Lift, Lift(X→Y))

[0130] = X를 포함하는 기사 중 Y를 포함하는 기사의 비율 / 전체기사 중 Y를 포함하는 기사의 비율

[0131] 예시적으로, 지지도는 추출된 기사 중 X와 Y를 모두 포함한 기사의 비율을 의미하며, 신뢰도는 X를 포함하는 기사 중 Y를 포함하는 기사의 비율을 의미한다. 또한, 향상도는 키워드를 입력했을 때 추출된 기사 중 Y를 포함한 기사 비율 대비 X를 포함하는 기사 중 Y를 포함하는 기사의 비율을 의미한다.

[0132] 자세히 말해 지지도는 X와 Y가 얼마나 많은 기사에서 동시에 사용되었는지를 측정해 준다. 만약 지지도 값이 매우 작다면 해당 연관규칙을 따르는 기사의 수가 매우 적다는 의미를 내포하게 된다. 즉, 지지도 측정을 통해 특정 단어집합이 특정기사에서만 사용된 것인지 아니면 많은 기사에서 사용된 것인지를 측정할 수 있다.

[0133] 신뢰도는 해당 규칙이 정답일 확률을 의미하며, 기사에서 X가 사용된 기사 중, Y가 동시에 사용된 기사의 비율을 의미한다. 만약 특정 단어 집합 간의 신뢰도가 매우 낮다면 해당 규칙이 정답이 아닐 확률이 높다는 것을 의미하며, 일반적인 경우에는 잘 발생하지 않는 규칙임을 알 수 있다. 지지도와 신뢰도가 모두 높은 연관성 규칙일지라도 유의미한 규칙이라 단정할 수는 없다. 왜냐하면 지지도와 신뢰도가 높은 연관규칙 중에는 우연에 의해 연관성이 높은 것처럼 보이는 규칙들이 있기 때문이다. 따라서 연관규칙의 유용성을 측정하기 위해 향상도가 사용된다.

[0134] 향상도는 'X를 포함할 때 Y를 포함할 확률이 X를 고려하지 않은 경우 Y를 포함할 확률에 비해 얼마나 향상되는가'에 대한 정보를 제공해준다. 만약 향상도가 1이면 X가 포함된 기사 중에서 Y를 찾는 것과, 모든 기사에서 Y

를 찾을 확률이 동일하기 때문에 두 단어집합 X, Y는 서로 상관관계가 없을 가능성이 높다. 반면, 향상도가 1 이상이면 두 단어집합은 서로 밀접한 상관관계를 (양의 상관관계) 가질 가능성이 높으며, 1 이하이면 서로 상반된 상관관계를 (음의 상관관계) 가질 가능성이 높다. 예를 들어, 빵, 버터는 밀접한 관계가 있는 양의 상관관계를 가질 확률이 높으며, 설사약, 변비약은 서로 상반된 관계가 있는 음의 상관관계를 가질 확률이 높다.

[0135] 단계 S901에서 유저 인터페이스부(110)는 요청 플래그 및 최소 빈도수를 전송할 수 있다. 데이터 분석부(130)는 유저 인터페이스부(110)로부터 요청 8(request 8)를 포함하는 요청 플래그를 수신할 수 있다. 요청8은 Word Cloud 시각화 요청을 목적으로하는 requestFlag 값일 수 있다.

[0136] 단계 S902에서, 분석부(133)는 분석 데이터 베이스(142)로부터 체언 집합 데이터를 수신할 수 있다.

[0137] 단계 S903에서, 분석부(133)는 체언 집합에 포함된 단어 중에서 최소 빈도수 이상의 단어를 결정하여 빈도수 분석을 수행할 수 있다. 또한, 분석부(133)는 결정된 최소 빈도수 이상의 단어를 빈도수에 따라 출력 위치, 출력 크기 및 출력 색을 결정하여 워드 클라우드(Word Cloud)형태로 출력할 수 있다.

[0138] 도 13은 본원의 일 실시예에 따른 결정된 최소 빈도수 이상의 단어를 빈도수에 따라 출력 위치, 출력 크기 및 출력 색을 결정하여 워드 클라우드 형태로 출력한 예를 나타낸 도면이다.

[0139] 예시적으로 도 13을 참조하면, 워드 클라우드 형태의 시각화하는 것은, 핵심적인 단어를 돋보이게 하는 시각화하여, 기사에서 사용된 단어들을 빈도수에 따라 서로 다른 크기와 색으로 표현할 수 있다.

[0140] 또한, 분석부(133)는 생성된 워드 클라우드 형태로 출력된 시각화 결과를 분석 데이터 베이스(142)에 저장할 수 있다.

[0141] 다음으로, 데이터 분석부(130)는 유저 인터페이스부(110)로부터 요청 9(request 9)를 포함하는 요청 플래그를 수신할 수 있다. 요청9은 Grouped Matrix 시각화 요청을 목적으로하는 requestFlag 값일 수 있다.

[0142] 단계 S904에서, 데이터 분석부(130)는 최소 지지도 및 최소 신뢰도 입력 정보를 수신할 수 있다. 또한, 데이터 분석부(130)는 제 2 단어 집합에 관한 입력을 수신할 수 있다.

[0143] 단계 S905에서, 분석 데이터 베이스(142)는 데이터 분석부(133)의 요청에 의해 매트릭스 데이터 셋트를 전송할 수 있다.

[0144] 단계 S906에서, 분석부(133)는 매트릭스 데이터 셋트를 기반으로 연관규칙 분석을 수행할 수 있다. 분석부(133)는 매트릭스 데이터 셋트 내의 빈발 단어들로 만들 수 있는 단어 조합을 포함하는 제 1 단어 집합을 결정하고, 제 1 단어 집합에 포함된 단어 조합이 기사에 포함되면 제 2 단어 집합에 포함된 단어도 기사에 포함되는지에 관한 연관규칙을 결정할 수 있다.

[0145] 분석부(133)는 결정된 연관규칙에 관한 지지도, 신뢰도 및 향상도를 연산하고, 결정된 연관규칙 중 최소 지지도 및 최소 신뢰도 이상의 지지도 및 신뢰도를 가지는 연관규칙에 관한 분석 결과를 출력할 수 있다.

[0146] 예시적으로 도14를 참조하면, 단계 S907에서, 분석부(133)는 제 1 단어 집합과 제 2 단어 집합 간의 연관성을 매트릭스 형태로 출력할 수 있다. 도 14는 그룹 매트릭스 출력의 예를 나타낸 도면이다. 도 14에서 연관규칙의 제 1단어집합(LHS)과 제2단어집합(RHS)을 기준으로 각 단어집합 간 연관성을 매트릭스(Matrix)형태로 나타낸 것을 확인할 수 있다. 그룹 매트릭스 출력 그래프의 원의 크기는 각 규칙의 지지도를 나타내고, 색상의 진함은 향상도를 의미할 수 있다. 제 1단어집합(LHS) 이름 앞의 숫자는 그 조건으로 되어있는 연관규칙의 수를 의미하며, 제 1단어집합(LHS)에 “+”와 함께 표시된 숫자는 표시가 생략된 단어의 수를 의미할 수 있다. 그룹 매트릭스 출력 그래프의 원의 크기가 클수록, 색이 진할수록 많이 발생하는 규칙이라고 해석할 수 있다.

[0147] 분석부(133)는 검색 키워드별로 서로 다른 개수의 규칙을 생성할 수 있다. 서로 다른 개수의 규칙이 생성되는 것은, 각 키워드마다 매트릭스 데이터 셋트에 포함되어 있는 기사의 수 및 기사 내 포함된 단어의 수가 다르기 때문이다. 분석부(133)는 옵션 값(지지도 =0.1, 신뢰도=0.1) 이상의 값을 가지는 규칙의 수가 검색 키워드별로 다르기 때문에, 검색 키워드별로 서로 다른 개수의 규칙을 생성할 수 있다. 다시 말해, 분석부(133)는 연관규칙 분석을 수행 시 각 기사에 포함되는 단어집합들에 따라 생성되는 규칙들의 지지도, 신뢰도 값이 결정할 수 있다. 매트릭스 데이터 셋트 내 기사의 수 및 각 기사에 포함되어 있는 단어의 수가 적을수록 옵션 값 미만의 규칙들이 많아져서 결과에서 제외될 수 있다.

[0148] 또한, 데이터 분석부(130)는 유저 인터페이스부(110)로부터 요청 10(request 10)를 포함하는 요청 플래그를 수신할 수 있다. 요청10은 Graph 시각화 요청을 목적으로하는 requestFlag 값일 수 있다.

[0149] 단계 S908에서, 데이터 분석부(130)는 분석 데이터 베이스(142)로 매트릭스 데이터 셋트를 요청할 수 있다. 분석부(133)는 매트릭스 데이터 셋트를 기반으로 연관규칙 분석을 수행할 수 있다. 분석부(133)는 특정 지지도 순위 이상의 규칙들을 기반으로 그래프(Graph) 시각화를 수행할 수 있다.

[0150] 단계 S909에서 연관 규칙 분석시 분석부(133)는 유저 인터페이스부(110)로부터 최소 지지도 및 최소 신뢰도값을 수신할 수 있다. 또한, 분석부(133)는 사용자가 원하는 단어에 대한 연관규칙 분석 결과만을 확인하기 위해 제2 단어집합(RHS)값을 설정할 수 있다.

[0151] 도 15를 참조하면, 단계 S910에서, 분석부(133)는 제 1 단어 집합에 속한 단어와 제 2 단어 집합에 속한 단어 간의 연관성을 네트워크 형태로 출력하는 그래프 출력을 할 수 있다. 도 15에 도시된 네트워트 그래프 형태의 단어에서 원으로 향하는 화살표는 제 1 단어 집합(LHS), 원에서 단어로 향하는 화살표는(RHS)를 의미한다. 예를 들어, 분석부(133)는 '송전탑(LHS)->O->주민(RHS)'의 연관성을 분석할 수 있다. 즉, 분석부(133)는 송전탑이라는 단어가 기사에서 사용된다면, 주민이라는 단어 또한 기사에서 사용된다는 규칙을 분석할 수 있다. 'O->전자파(RHS)'로 분석 될 경우, 제 1 단어 집합(LHS)에 제시된 단어의 사용여부에 상관없이, 전자파가 기사에서 사용된다는 규칙을 의미할 수 있다.

[0152] 또한, 원의 크기는 각 규칙의 지지도를, 원의 색상의 진하기는 향상도를 의미할 수 있다. 단어의 위치는 다른 단어들과의 연관관계를 의미하며, 도 15에서는 중계기가 연관관계의 중심에 있음을 확인할 수 있다. 즉, 원의 크기가 클수록, 색이 진할수록 발생 빈도가 높은 규칙이라고 해석할 수 있다. 일예로, 다른 단어들과 떨어져서 있는 두 단어가 있다면, 이는 두 단어가 둘만의 연관관계를 가지고 있는 것일 수 있다.

[0153] 분석부(133)는 생성된 네트워크 형태로 출력한 그래프 출력을 분석 데이터베이스(142)에 저장할 수 있다.

[0154] 또한, 데이터 분석부(130)는 유저 인터페이스부(110)로부터 요청 11(request 11)를 포함하는 요청 플래그를 수신할 수 있다. 요청11은 Scatter Plot 시각화(분포도 출력) 요청을 목적으로하는 requestFlag 값일 수 있다.

[0155] 단계 S911에서, 데이터 분석부(130)는 분석 데이터 베이스(142)로 매트릭스 데이터 셋트를 요청할 수 있다. 분석부(133)는 매트릭스 데이터 셋트를 기반으로 연관규칙 분석을 수행할 수 있다. 분석부(133)는 결정된 연관규칙의 분포를 분포도 출력으로 시각화하여 제공할 수 있다.

[0156] 단계 S912에서 연관 규칙 분석시 분석부(133)는 유저 인터페이스부(110)로부터 최소 지지도 및 최소 신뢰도값을 수신할 수 있다. 또한, 분석부(133)는 사용자가 원하는 단어에 대한 연관규칙 분석 결과만을 확인하기 위해 제2 단어집합(RHS)값을 설정할 수 있다.

[0157] 단계 S913에서 분석부(133)는 생성된 연관규칙들의 분포를 그래프로 제공할 수 있다. 예시적으로 도 16을 참조하면, 분도포 출력 그래프의 X축은 연관규칙의 지지도를 의미하고, Y축은 연관규칙의 신뢰도를 의미할 수 있다. 각 점들의 색상의 진하기는 향상도를 의미하는 것일 수 있다. 분포도 출력 그래프는 생성된 규칙들의 지지도, 신뢰도, 향상도에 대한 분포를 나타낼 수 있다. 분포도 출력 그래프의 각 점들은 하나의 규칙을 의미하며, 점들의 밀도는 규칙의 수에 비례할 수 있다. 분포도 출력 그래프는 검색 키워드에 상관없이, 지지도가 낮을수록, 신뢰도가 높을수록 생성된 규칙의 수가 많아지며, 규칙의 향상도는 높아질 수 있다. 연관규칙 분석에서, 발생 빈도가 낮은 규칙들 중 신뢰도가 높은 규칙일수록 두 단어집합 간 연관성이 높은 규칙임을 의미할 수 있다.

[0158] 마지막으로, 데이터 분석부(130)는 유저 인터페이스부(110)로부터 요청 12(request 12)를 포함하는 요청 플래그를 수신할 수 있다. 요청12는 Extraction Probability 시각화 요청을 목적으로하는 requestFlag 값일 수 있다.

[0159] 단계 S914에서, 분석부(133)는 분석 데이터 베이스(142)로 매트릭스 데이터 셋트를 요청할 수 있다. 분석부(133)는 매트릭스 데이터 셋트를 기반으로 빈도수 분석을 수행할 수 있다. 분석부(133)는 기사에서 특정 단어가 추출될 확률 분석 및 추출 빈도를 시각화하여 제공할 수 있다. 단계 S915에서, 분석부(133)는, 매트릭스 데이터 셋트를 읽어오고, 체언 집합에 포함된 단어가 매트릭스 데이터 셋트의 각 기사에서 추출될 확률을 연산하고, 체언 집합에 포함된 단어 별 각 기사에서의 추출 빈도 및 추출 확률을 포함하는 그래프를 출력하여 빈도수 분석을 수행할 수 있다.

[0160] 예시적으로 도 17을 참조하면, 단계 S915에서 분석부(133)는 추출 빈도 및 추출 확률을 포함하는 그래프를 바(Bar) 형태의 그래프로 표현할 수 있다. 도 17에 도시된 추출 빈도 및 추출 확률을 포함하는 그래프는 임의의 키워드를 입력했을 때 기사에서 특정 단어가 추출될 확률을 Bar 그래프 형태로 나타낼 수 있다. 추출 빈도 및 추출 확률을 포함하는 그래프의 X축은 빈도수가 높은 단어집합을 의미하며, Y축은 단어집합들의 추출확률을 의미한다. 도 17에 도시된 그래프에서는 추출된 모든 단어집합 중 빈도수가 높은 상위 30개의 단어집합이 X축을

구성한 예시를 나타내고 있다. 각 바 그래프 상단의 숫자는 단어집합 별 추출확률을 구체적을 나타낸 것이다.

[0161] 단계 S915에서 임의의 검색 키워드를 입력했을 때 특정 단어가 기사로부터 추출될 확률은 매트릭스 데이터 셋트를 기반으로 계산될 수 있다. 분석부(133)는 (특정 단어가 포함된 기사의 수) / (크롤링된 기사의 수)를 계산하여 추출 확률을 표현할 수 있다.

[0162] 분석부(133)는 생성한 추출 빈도 및 추출 확률 그래프를 분석 데이터 베이스(142)에 저장할 수 있다.

[0163] 상술한 설명에서, 각 단계는 본원의 구현예에 따라서, 추가적인 단계들로 더 분할되거나, 더 적은 단계들로 조합될 수 있다. 또한, 일부 단계는 필요에 따라 생략될 수도 있고, 단계 간의 순서가 변경될 수도 있다.

[0164] 본원의 일 실시 예에 따른 텍스트 데이터 수집 및 분석 방법은 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-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 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 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0165] 또한, 전술한 텍스트 데이터 수집 및 분석 방법은 기록 매체에 저장되는 컴퓨터에 의해 실행되는 컴퓨터 프로그램 또는 애플리케이션의 형태로도 구현될 수 있다.

[0166] 전술한 본원의 설명은 예시를 위한 것이며, 본원이 속하는 기술분야의 통상의 지식을 가진 자는 본원의 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.

[0167] 본원의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본원의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0168] 100: 텍스 데이터 수집 및 분석 장치

110: 유저 인터페이스부

120: 웹 크롤러부

121: 파싱부 122: 추출부

123: 언어 지원부 124: 파일 생성부

130: 데이터 분석부

131: 전처리부 132: 데이터 형성부

133: 분석부

140: 데이터베이스

141: 크롤링 데이터베이스 142: 분석 데이터베이스

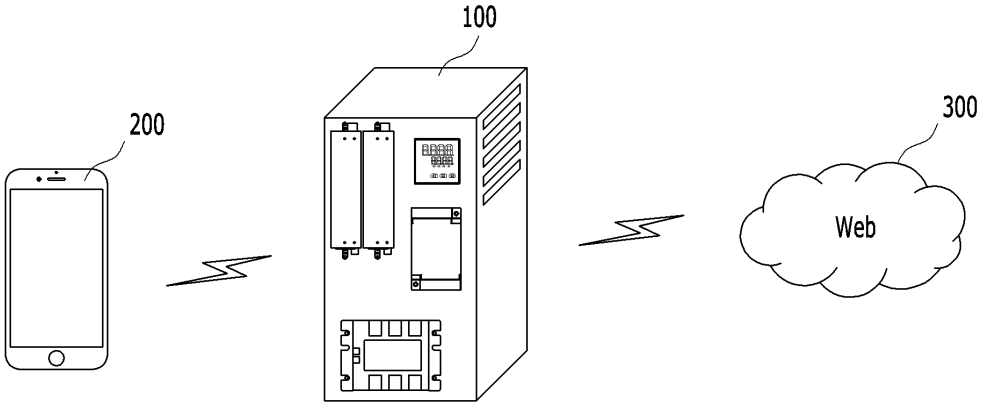
143: 사전 정의 단어 데이터베이스

200: 사용자 단말

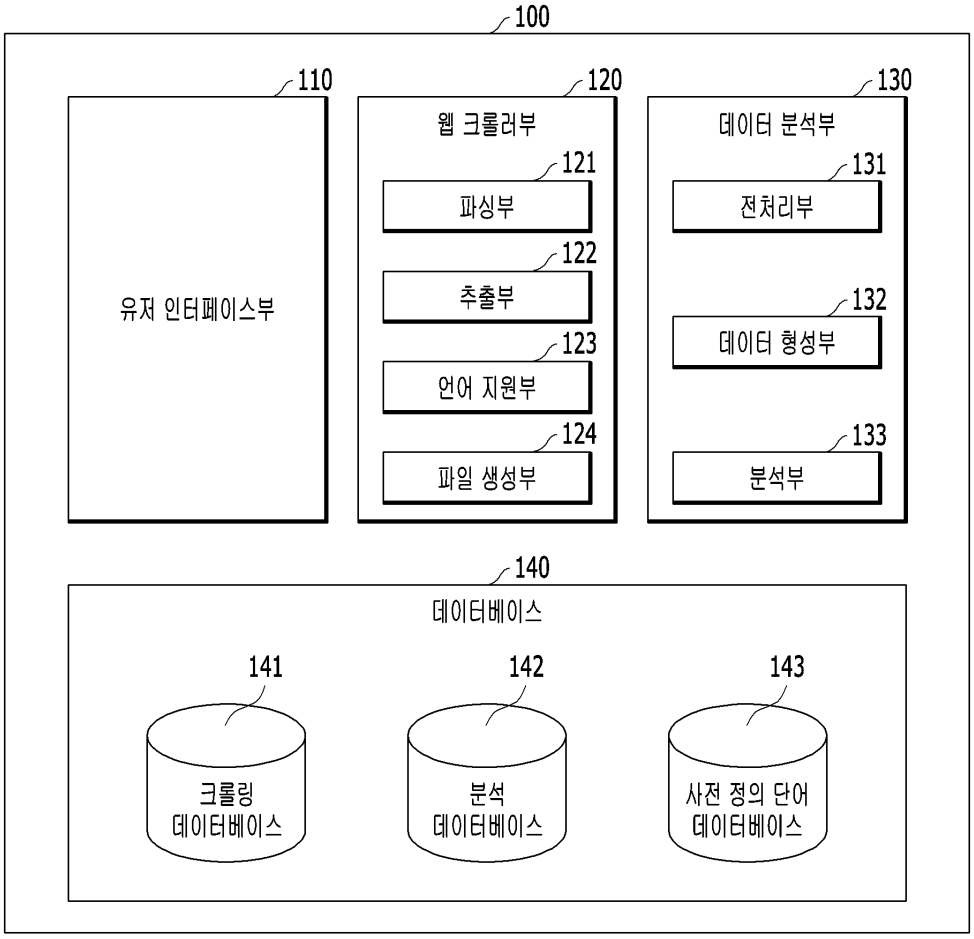
300: 웹

도면

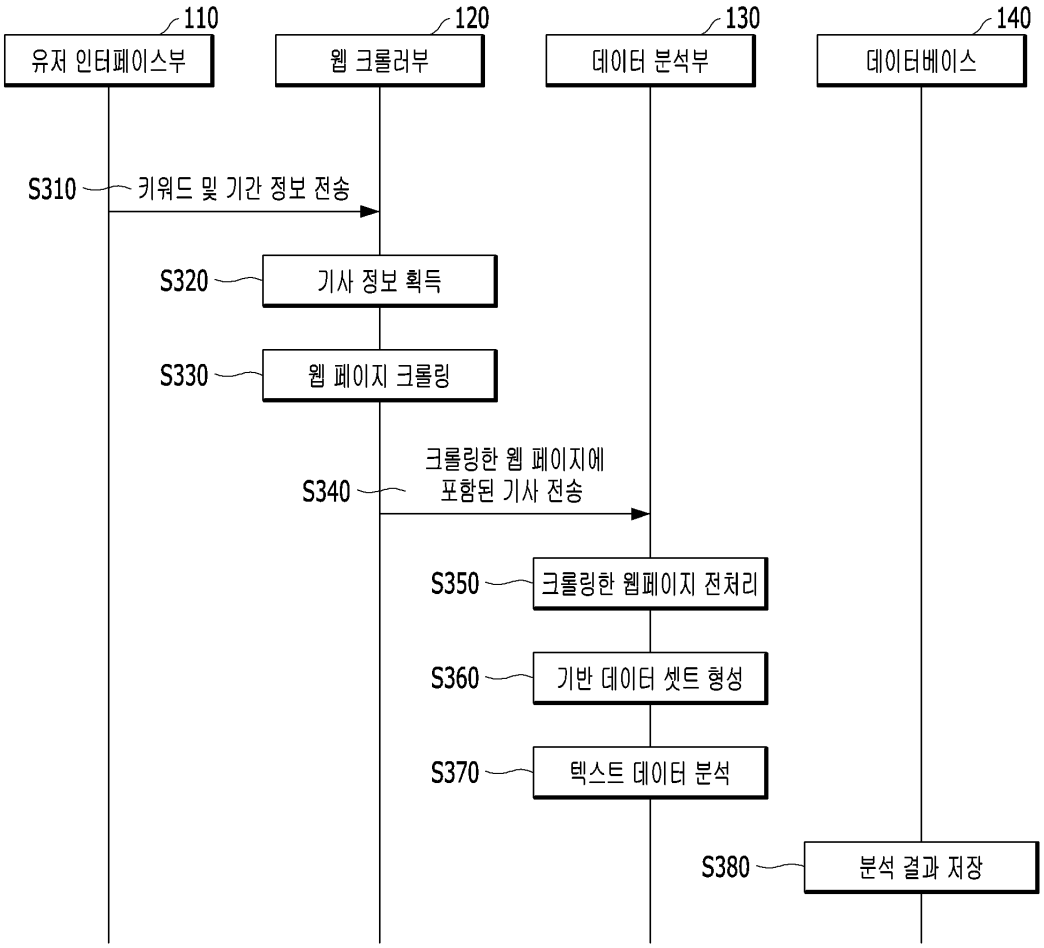
도면1



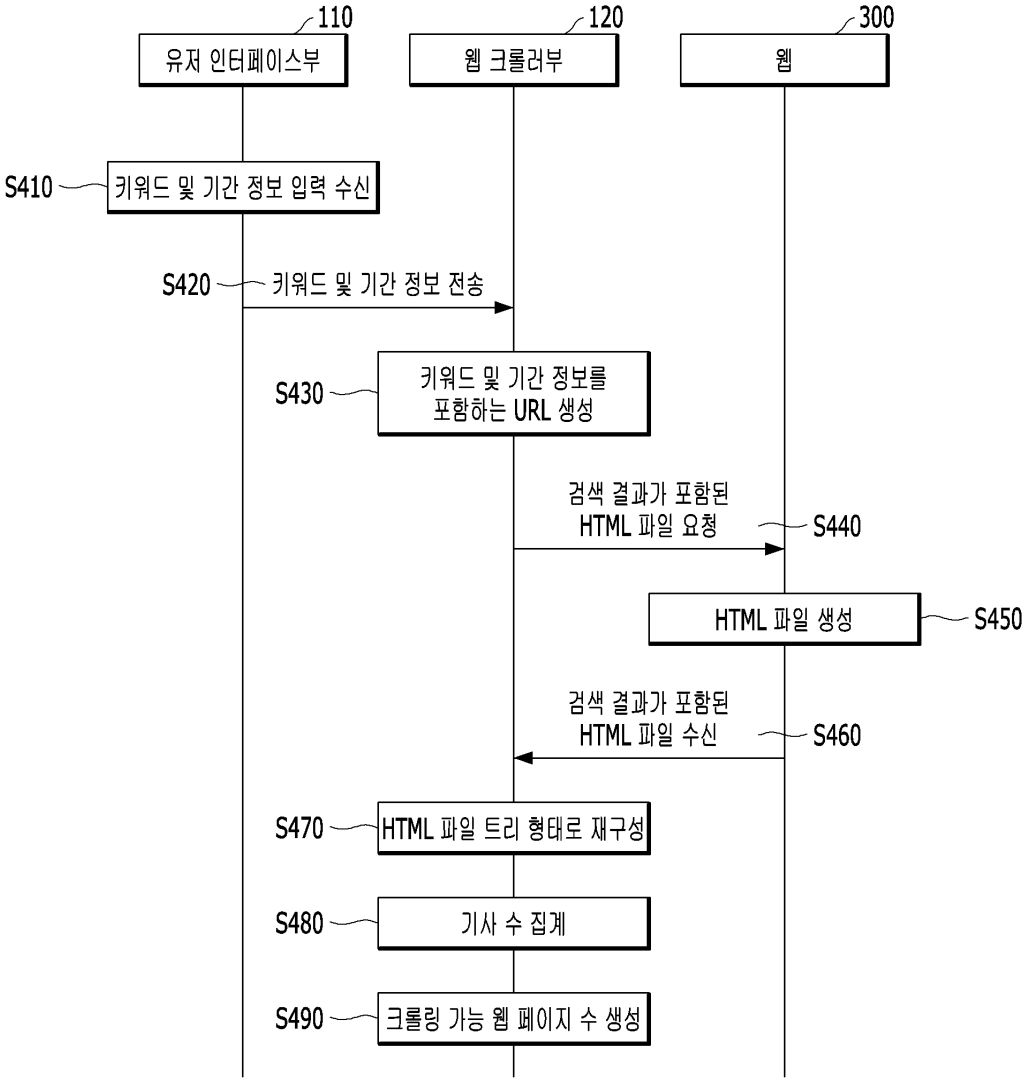
도면2



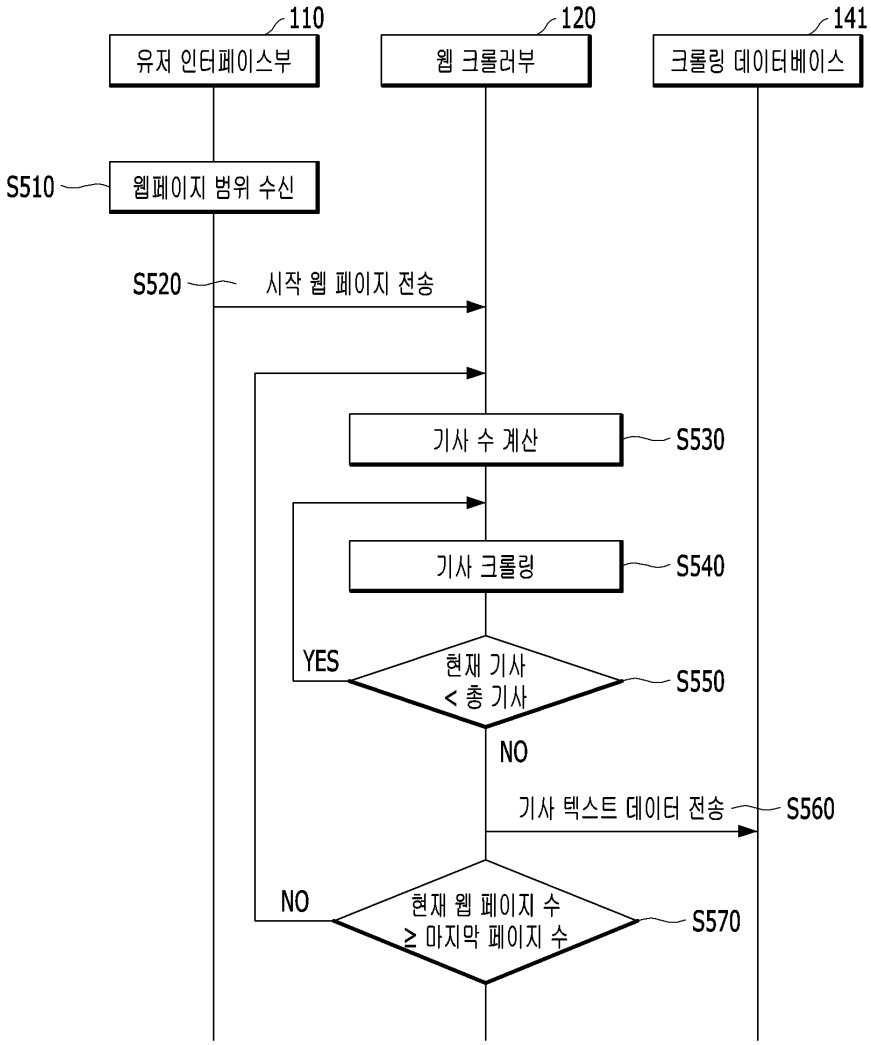
도면3


110
유저 인터페이스부
120
웹 크롤러부
130
데이터 분석부
140
데이터베이스
S310
키워드 및 기간 정보 전송
S320
기사 정보 획득
S330
웹 페이지 크롤링
S340
크롤링한 웹 페이지에
포함된 기사 전송
S350
크롤링한 웹페이지 전처리
S360
기반 데이터 셋트 형성
S370
텍스트 데이터 분석
S380
분석 결과 저장

도면4


110
유저 인터페이스부
120
웹 크롤러부
300
웹
S410
키워드 및 기간 정보 입력 수신
S420
키워드 및 기간 정보 전송
S430
키워드 및 기간 정보를
포함하는 URL 생성
검색 결과가 포함된
HTML 파일 요청
S440
HTML 파일 생성
S450
검색 결과가 포함된
HTML 파일 수신
S460
S470
HTML 파일 트리 형태로 재구성
S480
기사 수 집계
S490
크롤링 가능 웹 페이지 수 생성

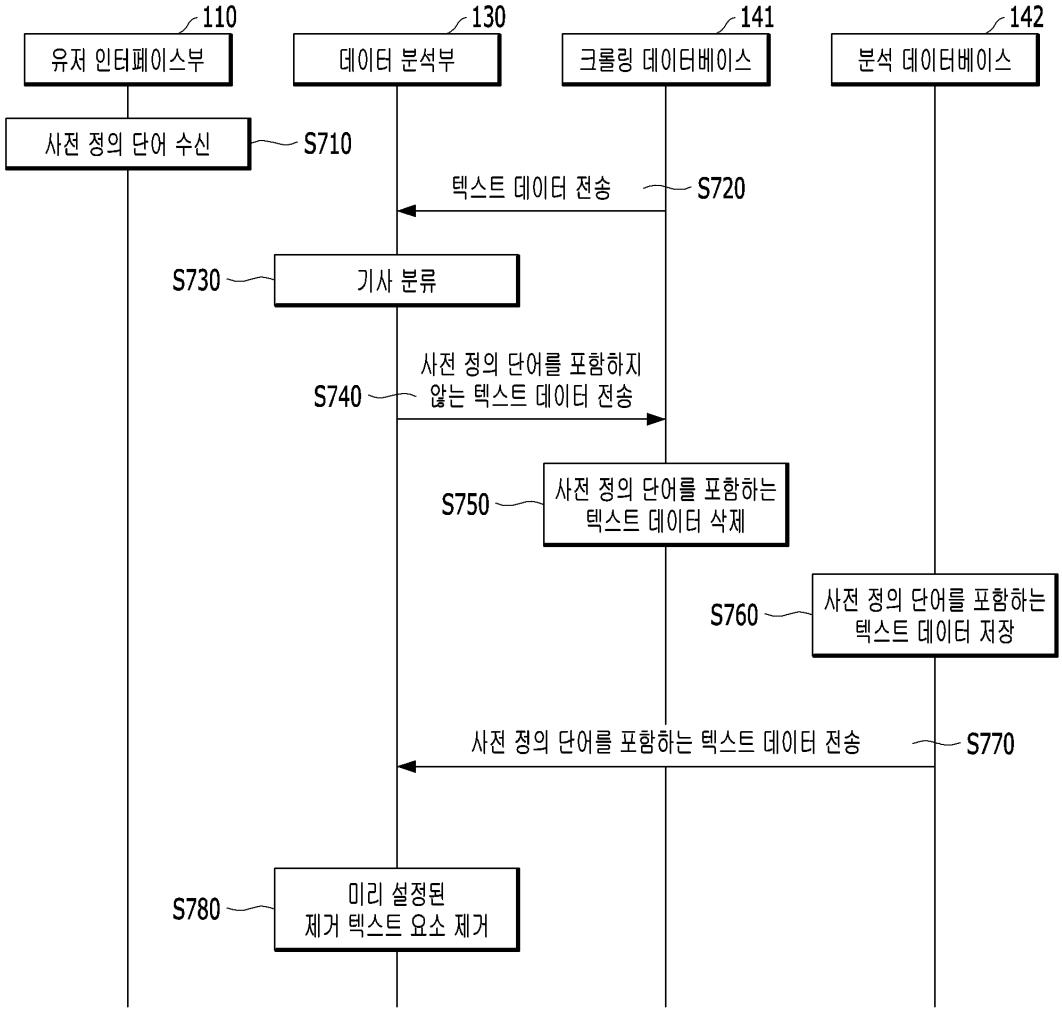
도면5



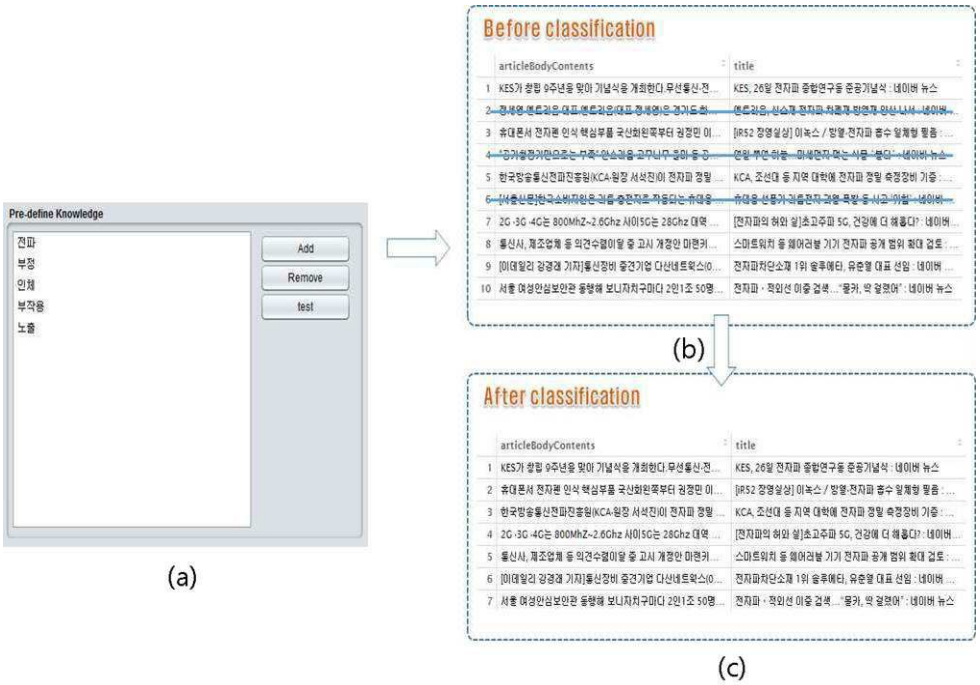
도면6

	articleBodyContents	title
1	KES가 창립 9주년을 맞아 기념식을 개최한다.무선통신·전...	KES, 26일 전자파 종합연구동 준공기념식 : 네이버 뉴스
2	정세영 엔트리움 대표.엔트리움(대표 정세영)은 경기도 화...	엔트리움, 신소재 전자파 차폐재·방열재 양산 나서 : 네이버 ...
3	휴대폰서 전자펜 인식 핵심부품 국산화왼쪽부터 권정민 이...	[iR52 장영실상] 이녹스 / 방열·전자파 흡수 일체형 필름 : ...
4	"공기청정기만으로는 부족" 안스리움·고무나무·율마 등 공...	연일 뿌연 하늘...미세먼지 먹는 식물 `불티` : 네이버 뉴스
5	한국방송통신전파진흥원(KCA·원장 서석진)이 전자파 정밀 ...	KCA, 조선대 등 지역 대학에 전자파 정밀 측정장비 기증 : ...
6	[서울신문]한국소비자원은 리튬 충전지로 작동되는 휴대용 ...	휴대용 선풍기 리튬전지 과열·폭발 등 사고 '위험' : 네이버 ...
7	2G ·3G ·4G는 800MhZ~2.6Ghz 사이5G는 28Ghz 대역 ...	[전자파의 허와 실]초고주파 5G, 건강에 더 해롭다? : 네이버...
8	통신사, 제조업체 등 의견수렴이달 중 고시 개정안 마련키...	스마트워치 등 웨어러블 기기 전자파 공개 범위 확대 검토 : ...
9	[이데일리 강경래 기자]통신장비 중견기업 다산네트웍스(0...	전자파차단소재 1위 솔루에타, 유춘열 대표 선임 : 네이버 ...
10	서울 여성안심보안관 동행해 보니자치구마다 2인1조 50명...	전자파 · 적외선 이중 검색..."몰카, 딱 걸렸어" : 네이버 뉴스

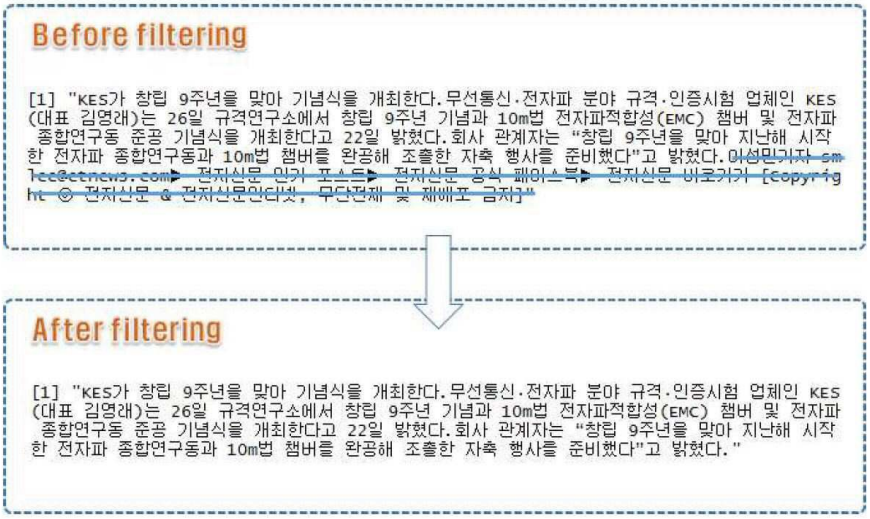
도면7


110
유저 인터페이스부
130
데이터 분석부
141
크롤링 데이터베이스
142
분석 데이터베이스
사전 정의 단어 수신
S710
텍스트 데이터 전송
S720
S730
기사 분류
사전 정의 단어를 포함하지 않는 텍스트 데이터 전송
S740
S750
사전 정의 단어를 포함하는 텍스트 데이터 삭제
S760
사전 정의 단어를 포함하는 텍스트 데이터 저장
사전 정의 단어를 포함하는 텍스트 데이터 전송
S770
S780
미리 설정된 제거 텍스트 요소 제거

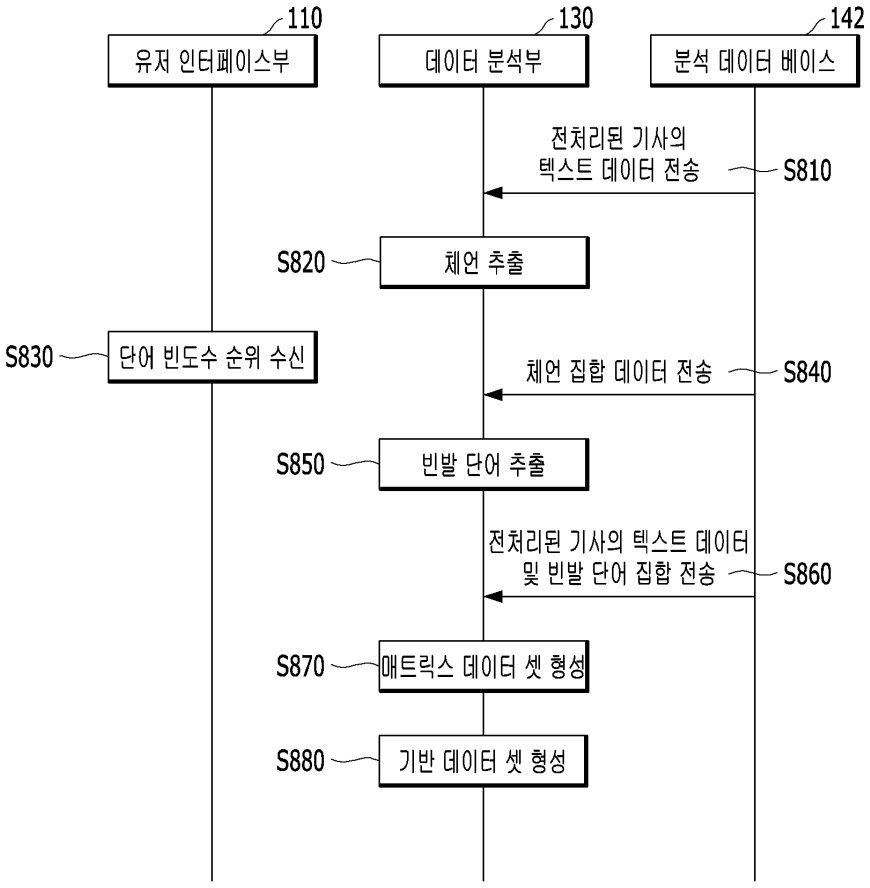
도면8


Pre-define Knowledge
전파
부정
인체
부작용
노출
Add
Remove
test
(a)
Before classification
articleBodyContents
title
(b)
After classification
articleBodyContents
title
(c)

도면9


Before filtering
[1] "KES가 창립 9주년을 맞아 기념식을 개최한다. 무선통신·전자파 분야 규격·인증시험 업체인 KES (대표 김영래)는 26일 규격연구소에서 창립 9주년 기념과 10m법 전자파적합성(EMC) 챔버 및 전자파 종합연구동 준공 기념식을 개최한다고 22일 밝혔다. 회사 관계자는 "창립 9주년을 맞아 지난해 시작한 전자파 종합연구동과 10m법 챔버를 완공해 조촐한 자축 행사를 준비했다"고 밝혔다. 이성민기자 smlee@etnews.com▶ 전자신문 인기 포스트▶ 전자신문 공식 페이스북▶ 전자신문 바로가기 [copyright © 전자신문 & 전자신문인터넷, 무단전재 및 재배포 금지]"
After filtering
[1] "KES가 창립 9주년을 맞아 기념식을 개최한다. 무선통신·전자파 분야 규격·인증시험 업체인 KES (대표 김영래)는 26일 규격연구소에서 창립 9주년 기념과 10m법 전자파적합성(EMC) 챔버 및 전자파 종합연구동 준공 기념식을 개최한다고 22일 밝혔다. 회사 관계자는 "창립 9주년을 맞아 지난해 시작한 전자파 종합연구동과 10m법 챔버를 완공해 조촐한 자축 행사를 준비했다"고 밝혔다."

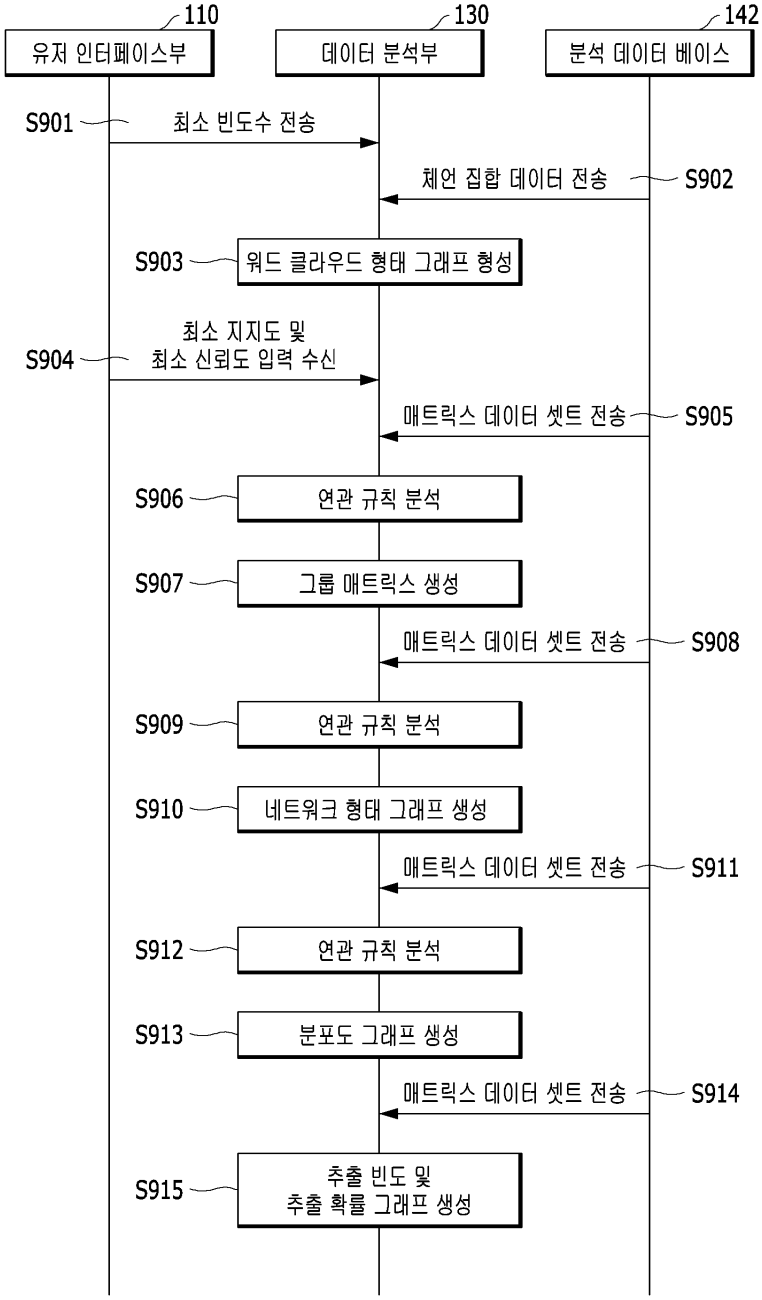
도면10


110
유저 인터페이스부
130
데이터 분석부
142
분석 데이터 베이스
전처리된 기사의
텍스트 데이터 전송
S810
S820
체언 추출
S830
단어 빈도수 순위 수신
체언 집합 데이터 전송
S840
S850
빈발 단어 추출
전처리된 기사의 텍스트 데이터
및 빈발 단어 집합 전송
S860
S870
매트릭스 데이터 셋 형성
S880
기반 데이터 셋 형성

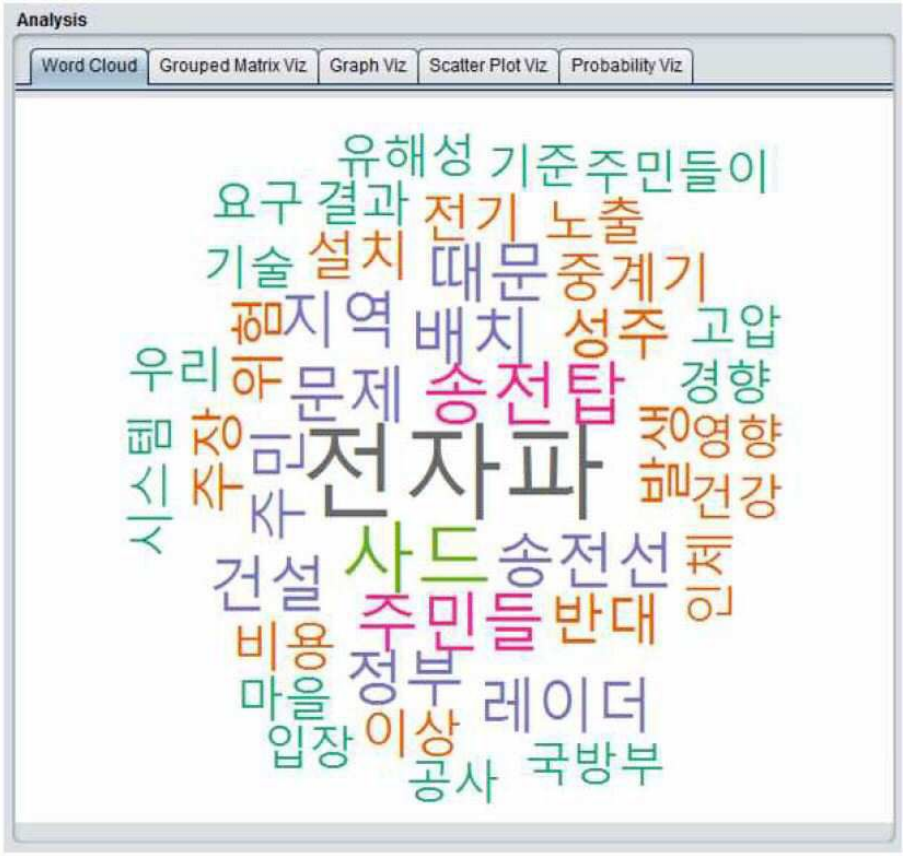
도면11

	전자파	레이더	사드	배치	영향	인체	국방부	정부	성주	문제	측정	지역	노출	때문	유해성	포대	결과	설치	경우	주파수
1	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1
2	1	0	0	0	1	1	0	1	0	0	1	0	1	0	0	0	0	0	1	0
3	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0
4	1	1	1	1	1	1	1	0	0	1	0	1	0	1	1	0	0	0	0	1
5	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0
6	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0
7	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	0
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
9	1	1	1	1	1	1	0	0	1	0	0	1	0	1	1	0	1	1	0	1
10	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	1	0

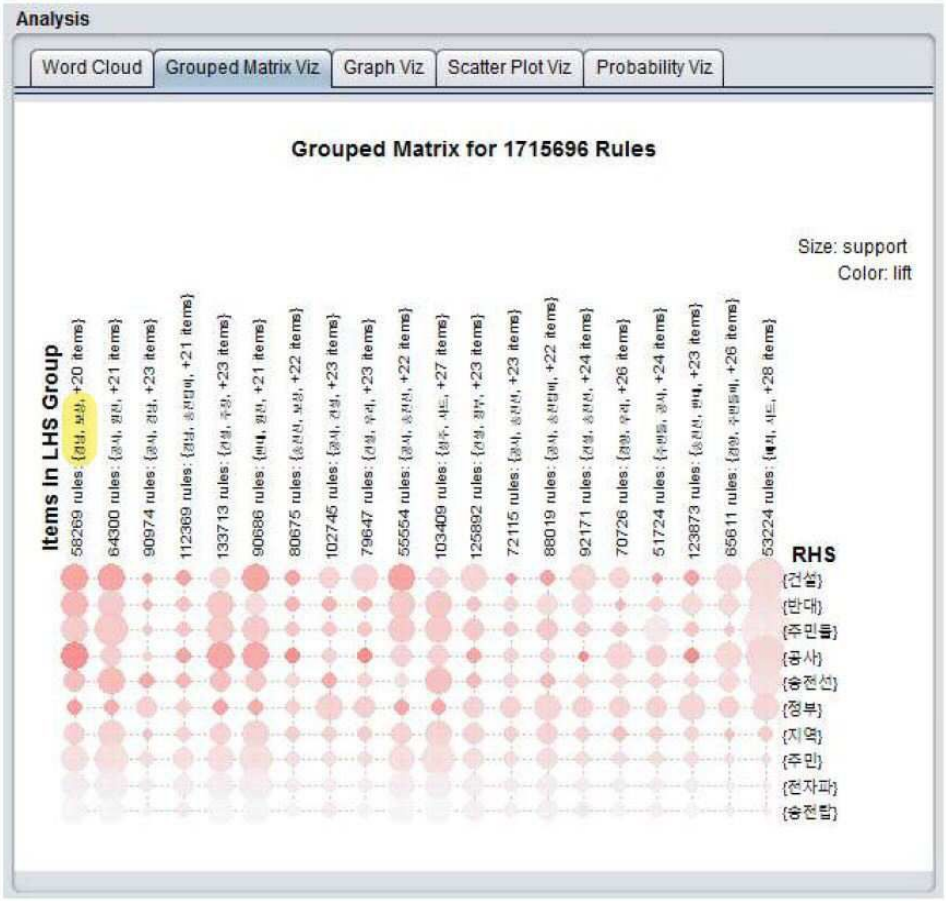
도면12


110
유저 인터페이스부
130
데이터 분석부
142
분석 데이터 베이스
S901
최소 빈도수 전송
체언 집합 데이터 전송
S902
S903
워드 클라우드 형태 그래프 형성
최소 지지도 및
최소 신뢰도 입력 수신
S904
매트릭스 데이터 셋트 전송
S905
S906
연관 규칙 분석
S907
그룹 매트릭스 생성
매트릭스 데이터 셋트 전송
S908
S909
연관 규칙 분석
S910
네트워크 형태 그래프 생성
매트릭스 데이터 셋트 전송
S911
S912
연관 규칙 분석
S913
분포도 그래프 생성
매트릭스 데이터 셋트 전송
S914
S915
추출 빈도 및
추출 확률 그래프 생성

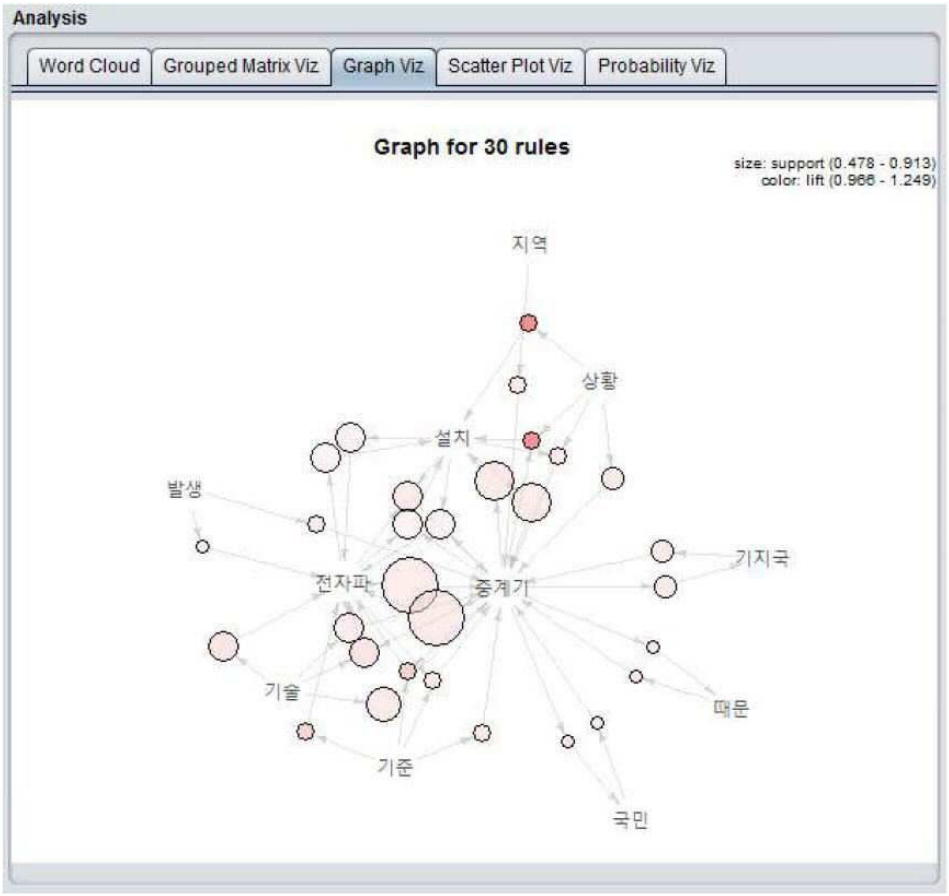
도면13


Analysis
Word Cloud
Grouped Matrix Viz
Graph Viz
Scatter Plot Viz
Probability Viz
유해성
기준
주민들이
요구
결과
전기
노출
기술
설치
때문
중계기
지역
배치
성주
고압
우리
문제
송전탑
경향
전자파
영향
건강
사드
송전선
건설
주민들
반대
비용
정부
마을
레이더
입장
이상
공사
국방부

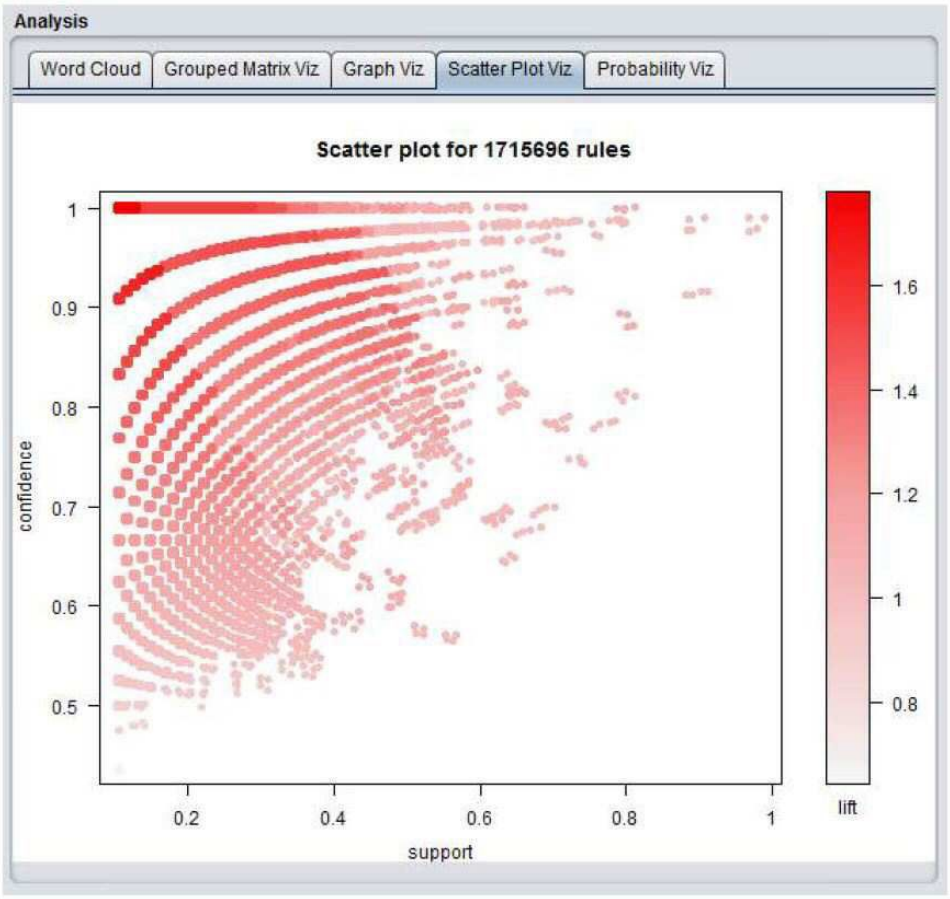
도면14


Analysis
Word Cloud
Grouped Matrix Viz
Graph Viz
Scatter Plot Viz
Probability Viz
Grouped Matrix for 1715696 Rules
Size: support
Color: lift
Items In LHS Group
58269 rules: {건설, 보상, +20 items}
84300 rules: {공사, 건설, +21 items}
90974 rules: {공사, 건설, +23 items}
112369 rules: {건설, 송전탑에, +21 items}
133713 rules: {건설, 주장, +23 items}
90686 rules: {반대, 건설, +21 items}
80675 rules: {송전선, 보상, +22 items}
102745 rules: {공사, 건설, +23 items}
79647 rules: {건설, 우려, +23 items}
55554 rules: {공사, 송전선, +22 items}
103409 rules: {청주, 시도, +27 items}
125892 rules: {건설, 정부, +23 items}
72115 rules: {공사, 송전선, +23 items}
88019 rules: {공사, 송전탑에, +22 items}
92171 rules: {건설, 송전선, +24 items}
70726 rules: {건설, 우려, +26 items}
51724 rules: {주변을, 공사, +24 items}
123873 rules: {송전선, 반대, +23 items}
85611 rules: {건설, 주변들에, +26 items}
53224 rules: {반대, 시도, +28 items}
RHS
{건설}
{반대}
{주민들}
{공사}
{송전선}
{정부}
{지역}
{주민}
{전자파}
{송전탑}

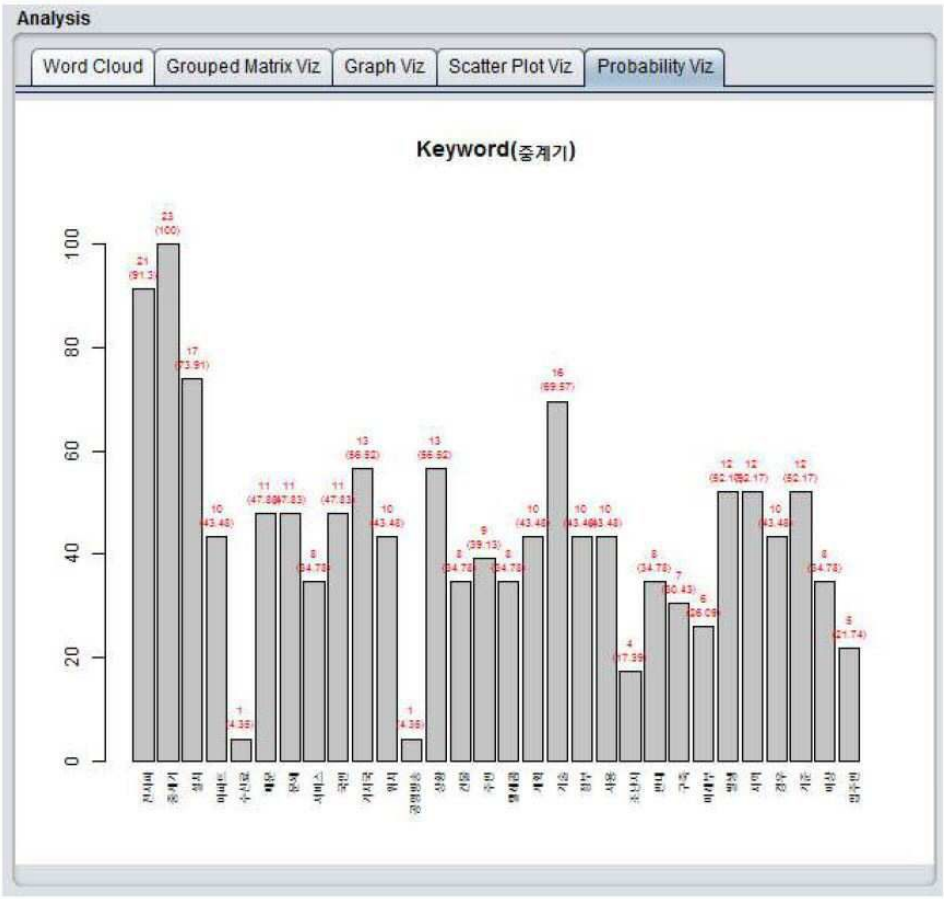
도면15


Analysis
Word Cloud
Grouped Matrix Viz
Graph Viz
Scatter Plot Viz
Probability Viz
Graph for 30 rules
size: support (0.478 - 0.913)
color: lift (0.966 - 1.249)
지역
상황
설치
발생
기지국
전자파
중계기
기술
때문
기준
국민

도면16


Analysis
Word Cloud
Grouped Matrix Viz
Graph Viz
Scatter Plot Viz
Probability Viz
Scatter plot for 1715696 rules
confidence
1
0.9
0.8
0.7
0.6
0.5
0.2
0.4
0.6
0.8
1
support
1.6
1.4
1.2
1
0.8
lift

도면17


Analysis
Word Cloud
Grouped Matrix Viz
Graph Viz
Scatter Plot Viz
Probability Viz
Keyword(중계기)
0
20
40
60
80
100

도면18

requestFlag 값	요청 목적
request1	검색된 기사의 수에 대한 요청
request2	기사 크롤링 요청
request3	데이터 분류 요청
request4	데이터 필터링 요청
request5	체언 추출 요청
request6	빈발단어 추출 요청
request7	매트릭스 데이터 셋 형성 요청
request8	Word Cloud 시각화 요청
request9	Grouped Matrix 시각화 요청
request10	Graph 시각화 요청
request11	Scatter Plot 시각화 요청
request12	Extraction Probability 시각화 요청