비정형 텍스트 분석을 위한 정보 추출 기술

KiSTi 한국과학기술정보연구원 Korea Institute of Science and Technology Information www.kisti.re.kr

특 허 명 비정형 텍스트 추출 성능 향상을 위한 시스템 및 방법

Keyword 빅데이터, 비정형데이터, 데이터마이닝, 웹 마이닝, 자연어처리

발 명 자 조민희

기 술 성

기술 개요

- 본 특허는 현실적인 문제에 나타나는 시간 또는 공간정보를 이용한 텍스트 정보 정확도와 추출 결과의 사실성을 판별하여 비정형 텍스트 추출 성능을 향상하기 위한 기술임

기존 기술 문제점

- 자연어의 다양한 표현과 사람들이 사용하는 은어 및 비유적 표현들의 텍스트 정보를 효율적으로 추출할 수 있는 기술이 있지만, 실제 현상을 반영한 사실을 추출할 수 있는 기술적 부재가 존재함
- 텍스트 정보를 추출하는 기술은 텍스트 자체에 포함된 정보로만 분석하고 있기 때문에 추출한 결과에 대한 신뢰도 여부와 결과 검증을 측정할 수 있는 기술적 부재가 존재함
- 기존의 비정형 텍스트 추출의 경우, 추출 정확도만 가지고 평가했기 때문에 소셜 네트워크(SNS) 등에 적용하면 잘못된 결과가 도출되는 문제가 발생됨

기술의 특징 및 우수성

▶ 기술의 특징

- 비정형 정보의 추출 데이터와 시계열 및 공간정보로 연계시켜 추출결과를 검증할 수 있음
- 이벤트 키워드(자연재해, 문서데이터 등)에 시간 또는 공간정보를 매핑 후 지식 후보들을 생성할 수 있음
- 비정형으로 추출한 데이터와 정형데이터를 시계열 및 공간정보로 연계시켜 취득 결과를 검증할 수 있음

▶ 기술의 우수성

- 기계학습 기법(SSVM)과 필터링 규칙을 단계적으로 적용한다면, 80.02%의 정보 추출 정확도 향상
- 분산·병력에 기초하여 지식 추출에 대한 성능을 개선한다면, 330KB/core&hour의 시맨틱(문장 해석) 지식 생성 속도를 향상할 수 있음
- 오픈 정보 기술 및 규칙을 적용한다면, 이벤트 정보의 추출 정확도 향상할 수 있음
- 추출 결과에 적합하지 않은 소셜 데이터를 제거한다면, 실제 상황에 맞는 정보를 취득하여 검증이 가능함

비정형 텍스트 분석을 위한 정보 추출 기술

상세설명

- 본 기술은 비정형 데이터 처리부, 정형 데이터 처리부, 필터부로 구성되어 있는 시스템임
- 뉴스, 블로그, 트위터 등으로부터 비정형 텍스트(미리 정의된 데이터 모델이 없는 정보)와 텍스트에 기재되어 있는 시간 등을 포함하는 수집상황 메타데이터(Metadata)를 수집함
- 이후 형태소(하나의 어절로부터 의미를 갖는 최소 단위) 분석 또는 개체명 인식(단어 유형)을 수행하여 오탈자 수정, 띄어쓰기 오류 등의 전처리를 하고 , 이벤트 키워드(자연재해, 문서데이터, 객체에 대한 사고 발생 경우 등)를 추출하고, 데이터 언어를 분석한 후 추출한 시간을 절대 정보로 변환하고, 공간 메타데이터를 이용하여 위치를 구체화함
- 구체화된 공간정보를 매핑시켜 추출 지식 후보들을 생성한 후, 시공간 연계 정형데이터로 추출 지식 후보들의 유효성을 판단하고, 그 판단 결과에 따라서 추출된 지식을 필터링함

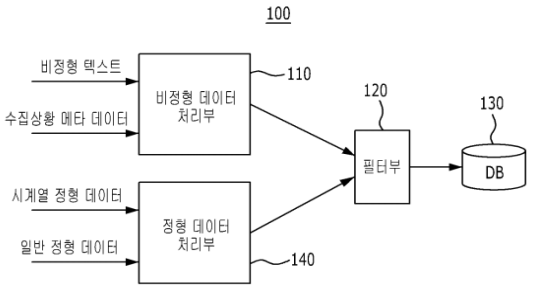

그림 1 비정형 텍스트 추출 성능 향상 시스템

기술완성도 (TRL)

기술완성도 : TRL5 (구현환경 적용실험)

TRL1	TRL2	TRL3	TRL4	TRL5	TRL6	TRL7	TRL8	TRL9
기술원리 발표	기술컨셉 설정	기술컨셉 증명	Lab Scale 시제품개발	**구현환경 적용실험**	Full Scale 시제품개발	유사 상용품 개발	상용품 완성	상용품 실시

활 용 분 야

활용분야 및 적용제품

활용분야

◆ 데이터 분석 분야
- 비정형 데이터 분석
- 실시간 데이터 분석
- 데이터 처리

◆ 데이터 관리 분야

◆ 데이터 플랫폼 분야

적용제품

◆ 텍스트 마이닝, 오피니언 마이닝 서비스
- 자연어 검색 시스템
- 자동 질의 응답 시스템
- 소셜 데이터 분석(구매 리뷰, 영화 리뷰 등)
- 사이버 범죄 예방
- 부정행위 탐지, 가짜뉴스 탐지

◆ 빅데이터 기반 마케팅 인텔리전스 서비스
- 텍스토미(패션 분석, 상권입지 분석 등)

비정형 텍스트 분석을 위한 정보 추출 기술

산업동향(기술 동향 및 트랜드 등)

- 현재 방대한 데이터가 생산되고 있으며 그중 비정형 데이터는 전체 데이터의 약 90% 정도 차지하고 있어 비정형 데이터의 분석에 관한 요구가 증가 하고 있음
- 텍스트 마이닝의 해외의 경우, 오피니언 마이닝, 소셜 네트워크 분석, 군집 분석 등 다양한 분석 분야에서 정확한 추출을 위한 연구가 진행중이고, 비정형 데이터를 원활히 처리할 수 있는 기술 중심으로 발전될 전망임
- 빅데이터 요소(양: Vlolume, 속도:Velocity, 다양성:Variety)의 분석을 통한 직관적인 정보제공을 위해 플랫폼의 역할이 중요시 되고 있으며, 데이터 추출과 분석을 위한 시계열 예측 모형(수요 예측 모델)과 머신러닝 기법의 활용이 요구되고 있음
- 최근 비정형 텍스트 추출 기술, 비정형 텍스트 데이터 처리 기술, 비정형 지수 산출 기술, 감성 및 사용자 사전 분류 기술, 자연어 처리 기술 등이 주목받고 있음

(출처2019년 중소기업 로드맵)

시장전망(목표시장 규모 및 전망)

- 시장조사기관인 IDC에 따르면, 빅데이터의 데이터 증가량은 매년 증가하고 있으며, 2020년 44제타바이트에 이를 것으로 전망했으며, 시장조사기관인 스태티스타의에 따르면, 2021년 64.5제타바이트에서 2025년 175제타바이트로 급성장할 것으로 예상됨
- 자연어 처리 기반 텍스트 마이닝의 세계 시장은 2017년 25,119백만달러에서 연평균 84.7%로 성장하여 2022년 540,000백만달러로 성장할 예정이며, 국내 자연어 처리(Natural Language Processing) 시장은 2020년 3,700억으로 전망되고 있음
- 또한, 자연어 처리 기술 중 자연어생성(사람의 언어를 계산 영역으로 가져오는 기술) 해외시장 규모는 2023년까지 8억 2530만 달러로 성장할 것으로 예상되며, 빅데이터 및 데이터분석 시장은 2018년 1,688억 달러에서 2022년 2,743억 달러까지 성장할 것으로 기대됨
- 미국 빅데이터 소프트웨어 시장은 2021년 240억 달러에서 연평균 20.7%로 증가하여 2027년 460억 달러로 예상되고 있음

(출처:TIPA_2019년 중소기업 로드맵, NIPA_글로벌 빅데이터 시장 보고서)

지재권현황

권리현황	특허등록번호	발명의 명칭
등록	10-1644429	비정형 텍스트 추출 성능 향상을 위한 시스템 및 방법

문 의 처

기술이전

한국과학기술정보연구원 Korea Institute of Science and Technology Information

담 당 자 심건욱 선임
연 락 처 042-869-0915
이 메 일 kwsim@kisti.re.kr

기술문의

한국과학기술정보연구원 Korea Institute of Science and Technology Information

담 당 자 조민희 연구원
연 락 처 042-869-0727
이 메 일 mini@kisti.re.kr