공학 석사학위논문

LDA 기반 데이터 모델링 기법을 이용한 SNS 여행지 추천

전남대학교 산업대학원
전기전자컴퓨터공학과

정 갑 수

2015년 8월

LDA 기반 데이터 모델링 기법을 이용한 SNS 여행지 추천

이 논문을 공학 석사학위 논문으로 제출함

전남대학교산업대학원
전기전자컴퓨터공학과

정 갑 수

지도교수 남 지 승

정갑수의 공학 석사 학위논문을 인준함

심사위원장 교 수 박 재 형 (인)
심 사 위 원 부교수 김 진 술 (인)
심 사 위 원 교 수 남 지 승 (인)

2015년 8월

목 차

표 목 차

그 림 목 차

LDA 기반 데이터 모델링 기법을 이용한 SNS 여행지 추천

정 갑 수

전남대학교 산업대학원
전기전자컴퓨터공학과 컴퓨터공학 전공
(지도교수 : 남지승)

(국문초록)

최근 들어 관광산업은 IT기술과 인터넷을 활용한 정보검색이 보편화 되면서 SNS를 이용한 관광정보 검색을 통해 여행지를 선택하고 여행상품을 구매하는 비중이 높아가고 있다. 더욱이 스마트폰의 보편화로 인해 언제 어디서나 정보를 손쉽게 획득할 수 있는 SNS의 특성에 의해 여행지 정보와 여행 체험 후기 등의 콘텐츠를 제공하고 있다.

실시간으로 소비자의 성향, 감정, 정서 패턴을 파악하여 개인 성향 맞춤형의 새로운 여행지 선별 서비스를 제공하는 방법으로 트위터, 블로그, 카페 등 SNS 상의 빅 데이터를 이용하는 것이 매우 효율적이다. 그러나 빅 데이터는 단지 그 막대한 정보의 양 뿐만 아니라 실시간 업데이트 되는 데이터의 생성 및 유통 속도와 매우 다양한 비정형, 비구조 데이터가 상호 융합되어 있다는 복합적인 특성을 가지므로 기존 데이터분석 기술로는 충분하지 않다.

최근 텍스트 뿐 아니라 이미지, 멀티미디어, 소셜 미디어 등의 분석에서도 널

리 쓰이고 있는, 대량의 텍스트 데이터에 내재되어 있는 시멘틱 패턴을 찾아내는 기법인 토픽모델링(Bayesian probabilistic topic modeling)을 고객 성향에 맞춘 여행지 선별을 위한 빅 데이터 분석 모델링에 사용한다. 토픽모델링 중 하나인 LDA(Latent Dirichlet allocation) 와 그의 추론 방법인 MCMC 샘플링 중심으로 토픽과 감성(sentiment)을 같이 다루는 기법을 제안하고자 한다.

본 논문에서는 빅 데이터 분석 플랫폼을 향후 다양한 분야에 적용하여 효율적인 빅 데이터의 분석과 비즈니스 활용 모델로서 기업 또는 국민 경제의 경쟁력 강화 및 생산성 향상 극대화를 도모하고자 한다.

1. 서 론

가. 연구 배경 및 목적

최근 스마트폰과 인터넷의 확산으로 실현된 정보화는 데이터 생산을 가속하고 있다. 인텔은 지난 2013년을 기준으로 인터넷에서 1분 동안 생산되는 데이터량을 연구·분석해 발표하였다. 1분 동안에 인터넷에서 일어나는 일은 유튜브에선 100시간 분량의 비디오가 업로드되며, 13만8889시간 분량의 비디오 시청이 이루어진다. 구글에선 410만건의 검색이 이루어진다. 페이스북에선 330만건의 콘텐츠가 공유되고, 690만개의 메시지가 전송된다. 트위터에선 34만7222건의 메시지가 '트윗'된다. 앱스토어에서는 19만 4064건의 앱 다운로드가 이루어진다. 아마존에서는 13만 34364달러의 거래가 이뤄진다. 판도라에선 3만 1773시간 분량의 음악이 재생된다. 인스타그램에서는 사진 3만 8194개가 업로드된다. 전세계적으로 전송되는 데이터량은 157만 2877GB(기가바이트)이다.

이와 같이 데이터 홍수와 폭증으로 대표되는 "빅 데이터"가 최근 특정 분야에 국한되지 않는 화두로 등장하였으며 빅 데이터 처리 및 분석 능력을 미래 경쟁력으로 인식하게 되었다. 세계경제포럼은 가장 주목할 기술로 빅 데이터를 지목하였으며 데이터 과잉 문제를 해결하고 데이터를 자산화 하여 활용하는 것을 최우선 현안으로 선정하였다.

본 연구는 세계 디지털 기술의 흐름에 발맞추어 빅 데이터 기반의 SNS 분석을 통한 사용자 성향에 따른 선별된 여행지 정보를 도출해내는 것을 중심으로 빅 데이터 처리 기술을 연구하고자 한다.

이를 통해, 방대한 데이터 쓰레기로 치부될 수 있는 사용자 클릭, 로그기록, 웹 페이지 방문횟수, 게시글 등의 데이터를 정보의 원천으로 삼아 과거 일률적인 여행서비스 형태를 탈피하고 소비자의 구매 성향, 감정, 정서 패턴을 파악

하여 개인 성향 맞춤형의 새로운 여행지 선별 서비스를 제공하는 빅 데이터 분석 플랫폼을 제시한다.

나. 연구의 범위

본 논문은 먼저 빅 데이터의 전망과 빅 데이터란 무엇인지에 대하여 알아본다. 빅 데이터의 국내외 연구 동향 및 기존 연구의 문제점으로 개발의 필요성을 나열하고, SNS 고객성향 분석을 통해 사용자의 감성적 측면을 반영한 여행지 선별 방법을 제안한다. 대량의 텍스트 데이터에 내재되어 있는 시멘틱 패턴을 찾아내는 기법인 토픽모델링 중 하나인 LDA(Latent Dirichlet allocation)와 그의 추론 방법인 MCMC 샘플링 기법을 중심으로 토픽과 감성(sentiment)을 같이 다루는 알고리즘을 제안 했을 때 학술적, 기술적 측면의 활용방안 및 기대성과 제안으로 본 논문을 마무리 한다.

2. 관련 연구

2.1 빅 데이터란

디지털 경제의 확산으로 우리 주변에서는 규모를 가늠할 수 없을 정도로 많은 정보와 데이터가 생산되는 '빅 데이터(Big Data)' 환경이 도래되고 있다.

빅 데이터는 말 그대로 큰(Big) 자료(Data)를 의미한다고 볼 수 있는데 디지털 환경에서 생성되는 데이터로 그 규모가 방대하고, 생성 주기도 짧고, 형태도 수치 데이터뿐 아니라 문자와 영상 데이터를 포함하는 대규모 데이터를 말한다.

과거 인터넷상에서 생산된 게시물, 이용자의 클릭 및 검색기록, 로그, 각종 신호등의 데이터 개념은 단순한 저장으로 구조화되고 용량만 차지하는 의미없는 데이터로 치부되었다. 그러나 최근 데이터의 영역은 각종 스마트기기가 본격적으로 보급되고 수집된 데이터 속에서 가치 있는 정보를 찾아 개인과 조직의 형태를 추론, 여론을 파악하거나 개인 생활 패턴의 흔적 등 비구조화된 데이터로 양산되고 있으며, 이는 정치, 사회, 경제, 문화, 과학 기술 등 전 영역에서 활용가능성을 모색하여 새로운 비즈니스의 가치 있는 자산이 되고 있다.

가. 빅 데이터의 특징

빅 데이터의 특징은 많은 기관에서 공통적으로 페타(Peta: 1015), 엑타(Exa: 108), 제타(Zeta: 1021)바이트 등 기존의 데이터 단위를 넘어서는 엄청난 데이터의 양(Volume), 데이터의 생성과 흐름이 매우 빠르게 진행되는 속도(Velocity), 사진, 동영상 등 기존의 구조화된 데이터가 아닌 형태의 다양성(Variety)을 기본으로 가치(Value)나 복잡성(Complexity)의 특성이 있다.

이는 스마트폰 및 소셜 미디어의 일반화로 규모(Volume)이 증가하고 실시간

데이터 축적 및 분석으로 속도(Velocity)가 높아지고 뉴스댓글, 전기신호, 음성, 기상 데이터등 다양한 소셜 미디어등의 다양한 데이터(Variety)가 증가되었으며 의도되지 않은 비정형화되고 혼잡하게 분산된 데이터와 목적간의 관계가 복잡(Complexity)해졌다.

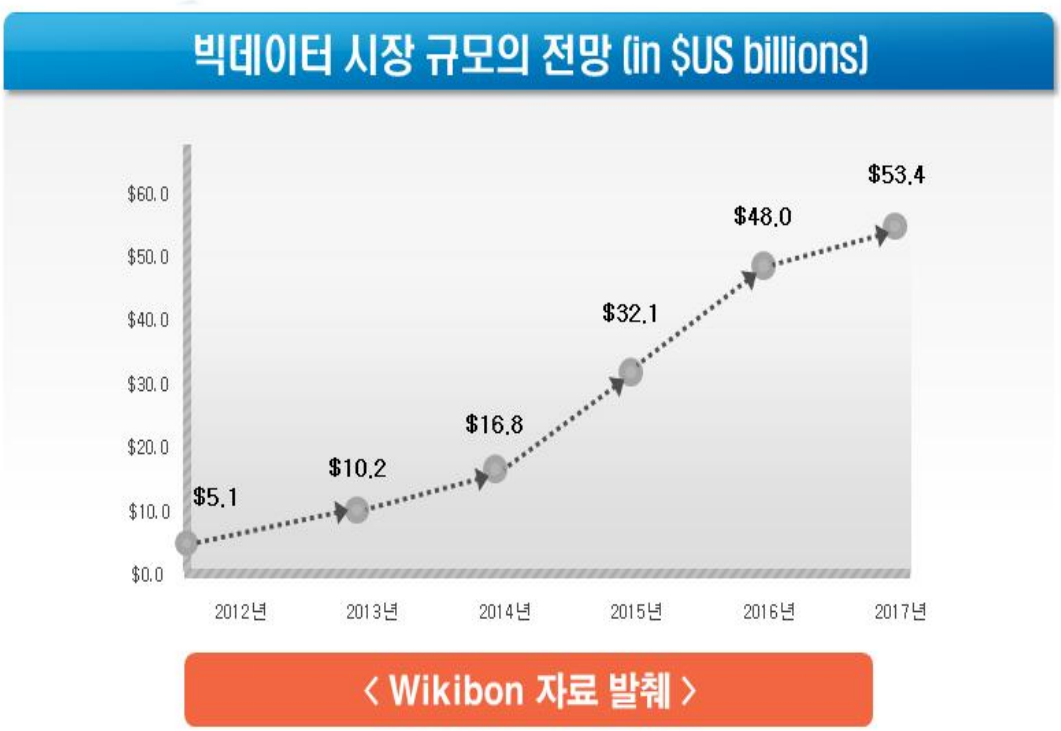
구분	기존	빅 데이터 환경
데이터	- 정형화된 수치자료 중심	- 비정형의 다양한 데이터 - 문자 데이터(SMS, 검색어) - 영상 데이터(CCTV, 동영상), 위치 데이터
하드웨어	- 고가의 저장장치 - DataBase, Data-warehouse	- 클라우드 컴퓨팅 등 비용 효율적인 장비 활용 가능
소프트웨어/분석 방법	- 관계형 데이터베이스(RDBMS) - 통계패키지(SAS, SPSS) - 데이터 마이닝(data mining) - machine learning, knowledge discovery	- Hadoop, NoSQL - 오픈 소스 통계솔루션(R) - 텍스트 마이닝(text mining) - 온라인 버즈 분석(opinion mining) - 감성 분석(sentiment analysis)

<표2-1> 빅 데이터 환경의 특징

나. 빅 데이터의 전망

빅 데이터를 통해 경쟁에서 앞서 나가려는 기업들의 욕구, 그리고 클라우드의 영향으로 인해 IT 투자가 감소할 것으로 예상되는 상황에서 새로운 시장을 발굴하려는 IT서비스 기업들의 욕구가 빅 데이터의 앞날에 긍정적인 영향을 미치고 있다. 그에 따라 빅 데이터 시장 규모는 대형 ICT 업체들과 신생업체들 간의 고객 및 시장점유율 경쟁이 가속화되면서 시장이 빠르게 확대되어 가고 있다. 향후 5년간 지속적으로 확대될 예정이며, 2017년경에는 530억 달러

를 돌파할 것으로 전망되고 있다.



<그림2-1> 빅 데이터 시장 규모 전망 (출처 : 위키본)

이와 같이 빅 데이터가 차세대 ICT를 이끌어 갈 핵심동력으로 주목받는 가장 주된 요인은 기존과 차별화된 대용량 데이터의 새로운 분석과 전망을 통해 새로운 서비스를 개발할 수 있는 가능성 때문이다.

특성	효과
대규모 (large scale)	- 기술 발전으로 데이터를 수집, 저장, 처리 능력 향상 - 현실세계 데이터를 기반으로 한 정교한 패턴분석 가능 - 데이터가 많을수록 유용한 데이터, 전혀 새로운 패턴의 정보를 찾아낼 수 있는 확률도 증가
현실성 (Reality)	- 우리사회 일상에서의 데이터 기록물의 증가 등 현실 정보, 실시간 정보의 축적이 급증될 전망 - 개인의 경험, 인식, 선호 등 인지적인 정보 유통 증가
시계열성 (Trend)	- 현시점뿐만 아니라 과거 데이터의 유지로 시계열적 연속성을 갖는 데이터의 구성 - 과거, 현재, 미래 등 시간 흐름상의 추세분석 가능
결합성 (Combination)	- 의료, 범죄, 환경 등 타 분야, 이종 데이터 간의 결합으로 새로운 의미의 정보 발견 - 실제 물리적인 결합 이전에, 데이터 결합을 통한 사전 시뮬레이션, 안정성 검증 분야 발전 가능

<표2-2> 빅 데이터 파급효과

2.2 빅 데이터를 활용한 고객 성향 맞춤 여행지 선별

1980년대 후반 경제성장과 더불어 라이프스타일 및 생활수준이 높아지고 관광을 즐길 수 있는 여가 시간이 풍부해 짐으로써 여행의 이슈는 앞으로도 계속 되는 윤택한 삶의 주제가 될 것이다. 또한 해외여행에 대한 기대 또한 커지고 여러 가지 여행에 대한 상품 또한 빠른 성장을 보이고 있다. 여행은 가보지 못한 곳의 동경이기도 하지만 미리 소비자의 취향 및 감정, 정서를 미리 파악하여 추천된 여행지라면 그것이야 말로 금상첨화 일 것이다. 그리하여 이러한 가치창출의 기회를 이용할 수 있는 빅 데이터를 이용하여 단순한 여행의 상품보다는 다양한 여행의 패턴을 만들어 여행지에서의 색다른 스타일의 문화즐기기 , 개인성향에 맞는 휴양스타일, 쇼핑을 추구하는 여행에 이르기까지 선별되어 질 수 있을 것이다.

인터넷 여행 쇼핑몰의 예를 들어보자. 고객이 원하는 여행지를 선택하고 상품을 구매하게 되면 과거에는 여행사에서 그 정보만 기록되었지만 지금은 구매를 하지 않더라도 고객이 정보를 보고 사용했던 기록이 자동적으로 데이터로 저장되며 기존에 구매 이력 정보를 통해 어떤 여행상품을 구매했는지 그 데이터를 분석하고 또한 관련된 같은 취향의 소비패턴 고객들이 구매했던 다른 여행지의 상품도 추천하여 보여준다. 또한 특정한 사이트뿐만 아니라 트위터, 블로그나 SNS에서 유통되어지는 텍스트 정보의 내용을 통해서 생활 패턴, 습관, 성향뿐 아니라, 소통하는 상대방의 연결 관계 스토리 까지도 빠른 시간내 분석이 가능하여 보다 나은 많은 여행 정보를 제공해 줄 수 있는 것이다. 이러한 제공 형태는 고객의 데이터 분석을 통해 현 고객의 유지와 향후 이탈 방지등 향후 마케팅 까지 진행 할 수 도 있다.

고객의 데이터 분석은 빅 데이터 시대를 맞이하여 고객이 여행 상품에 대해 관심을 갖게 되고 고객의 욕구에 맞는 여행상품을 제공하기 위해, 고객이 인지하는 부분과 그들의 요구에 따라 상품을 개발할 수 있는 자료를 제공함으로써 보다 삶의 질을 향상시키고 윤택한 여행 및 선택된 장소를 방문하게 될 것이다.

3. 국내외 연구 동향 및 기존 연구의 문제점

3.1 국내외 연구동향

가. 공공 행정 분야 활용 사례

세계 각국의 정부들은 빅 데이터가 미래사회를 위한 새로운 경제적 가치의 원천이 될 것으로 기대하고, 의료나 소매, 제조, 개인위치 정보 이외의 각 분야에 걸쳐 데이터 분석에 의한 예측과 서비스 개발을 위해 노력하고 있다. 데이터 분석을 통해 사회현상에 대한 미래 통찰력을 얻을 수 있으며, 위험 징후 및 이상신호를 포착할 수 있기 때문이다.

국가	추진기관	세부 내용
미국	대통령 직속 과학기술정책실	- '빅 데이터 연구개발 이니셔티브' 발표(2012. 3) - 부처별, 지방정부별 빅 데이터 활용한 서비스 발굴 및 운영 - 공공정보 데이터 공개사이트 'data.gov' 운영
영국	기업혁신기술부	- '데이터 전략위원회(Data Strategy Board)' 설립(2012. 3) - 데이터 공유플랫폼 'data.gov.uk' 운영
싱가포르	경제개발청	- 국가안보조정국(NSCS) 내 RASH 시스템 마련(2004. 7) - 민간협력으로 데이터분석연구소 설립 - 공공정보 데이터 공개사이트 'data.gov.sg' 운영
대한민국	대통령 직속 국가정보화전략위	- '빅 데이터를 활용한 스마트정부 구현방안' 마련(2011. 10) - 빅 데이터 마스터플랜 추진 및 빅 데이터 국가전략포럼 실시 - 공유자원포털 'data.gov.kr' 운영

<표3-1> 공공 행정 분야 활용 사례

나. 기업 비즈니스 분야 활용 사례

빅 데이터 현상은 거의 모든 비즈니스 부분에서 진행되고 있다. 미국의 경우, 현재 금융서비스, 통신, 전력, 미디어, 소매, 에너지등 거의 모든 산업 분야에서 빅 데이터 기술을 도입중인 반면에 국내 기업은 아직 미비한 단계이다. 모든 기업이 보유한 데이터가 핵심 가치를 알아낼 수 있을 만큼 충분한 양에 도달해 어떤 기업이 먼저 그 가치를 창출하느냐가 향후 기업의 성패와 직결되는 상황에 직면해 있다.

산업	국가	기업	특 징
제조	미국	포드	자사 차량의 내부통신망에서 수집한 정보를 스마트폰에서 이용 가능한 'OpenXC' 프로젝트 추진
	일본	히타치플랜트테크놀로지	크레인 곳곳에 센서를 부착해 무게중심 이탈 여부, 오작동 징후를 파악하는 '크레인 닥터 클라우드' 시스템구축
	미국	마이크론테크놀로지	제품 생산시간에 영향을 미치는 요소를 분석해 비용절감 방안을 수립
	한국	포스코	SAS를 통해 생산고정별 온도, 습도, 압력, 성분 등의 비철강 생산 전 과정의 각종 데이터를 0.001초 단위로 수집. 분석함으로써 불량률을 최소화, 생산효율성 높이기 위한 실시간 공정제어
유통	미국	월마트	판매 지역에 따른 고객의 선호도 파악 및 수요에 따라 물류 조절 및 재고 예측
	스페인	자라	전 세계 매장의 판매 데이터를 분석을 통해 글로벌 트렌드 실시간 탐지하고, 상품 유통망 프로세스 개선을 통해 무재고 시스템 실현
	미국	이베이	고객의 SNS 활동내용과 과거 구매이력을 분석하여 명절·기념일에 선물리스트 작성 및 추천
서비스	미국	구글	방문자의 검색어를 바탕으로 일상생활과 밀접한 각종 정보 제공 및 안드로이드 디바이스를 통한 사용자 정보보유, 빅 데이터 분석, 분산처리 수행, 클라우드 컴퓨팅
	미국	페이스북	실시간 입찰 광고 플랫폼 'FBX'을 통해 이용자들의 정보와 검색어를 실시간으로 분석하여 맞춤형 광고 제작, 방대한 소셜데이터 보유

서비스	미국	애플	아이튠즈 스토어, 아이클라우드 서비스를 제공, 이용자들의 데이터 수집, 분석, 음성인식 서비스 '시리(siri)'이용 질문이나 행동을 미리 예측해 최적의 답을 제시
	미국	넷플릭스	고객이 대여한 영화목록 등을 분석해 개인별 맞춤형영화 콘텐츠 제공
	한국	SK 텔레콤	네비게이션 서비스 티맵을 통한 전국 도로의 교통상황, 길안내 도착시간,GPS 장치를 통한 전국도로의 교통 정보 수집

<표3-2> 기업 비즈니스 분야 활용 사례

주요국 정부는 정부 데이터를 공개하는 전용사이트를 만들어서 데이터를 활용하고 새로운 지식을 만들기 위해 노력 하고 있다. 공공 부문의 데이터 공개를 통해서 정부의 투명성을 높이고, 국민의 알 권리를 향상 시키며, 시간과 자원의 절감 효과를 지향하고 있다.

3.2 기존 연구의 문제점

현재 공공 . 행정 분야 외의 기업 비즈니스 분야에서는 전략적으로 빅 데이터를 활용 하여 높은 가치를 창출하고 있다. 반면, 국내 기업은 빅 데이터에 대한 인식이 부족하여 정보를 체계적으로 축적하지 못했고, 단지 의사결정합리화를 위해 일회적으로 데이터를 소모하는 경우가 많아 빅 데이터 관리와 분석에 필요한 지식기반이 취약한 실정이다. 빅 데이터 관리와 분석에 필요한 지식기반 취약과 인프라 부족 등으로 빅 데이터를 경영에 접목시킨 사례가 많지 않으며, 주로 글로벌 기업의 기술과 오픈소스 솔루션에 의존하고 있다.

기본적인 수준에 머물러 있던 기계학습 관련 기법들은, 그것의 잠재력을 최대치로 끌어낼 수 있는 빅 데이터의 출현으로 인해 점점 실현 가능성이 높아지고 있다. 빅 데이터를 처리하기 위해서는 기존의 기계학습 방법은 대규모성(scalability)을 갖고 있지 못하기 때문에, 병렬처리 기법을 이용한 접근 방법을

많이 사용하였다. 그러나 이 경우는 기계학습 방법 그 자체를 바꾸기보다는 연산 능력과 저장 공간을 병렬적으로 처리하는 것으로 확장에 한계가 존재한다. 이를 극복하기 위해서 기계학습 기법 자체를 개선하여 빅 데이터 시대에 대응할 수 있게 하는 움직임이 병렬처리 방법 기반 위에서 빅 데이터를 처리할 수 있는 연구가 진행되고 있다.

기계학습 중 감독학습(supervised learning)을 통해 학습 데이터를 구축하는 것은 시간과 비용이 많이 들고, 이렇게 구축된 학습 데이터도 적용 분야(domain)가 변경될 경우 다시 재구축 비용이 발생하게 된다. 결국 학습 데이터 적용 분야가 변경될 때마다 각각의 분류기를 구축해야 하는데, 이것은 현실적으로 불가능하다. 이런 문제를 도메인 적응 문제라고 정의하는데, 이를 해결하기 위한 여러 가지 방법이 연구 되고 있다.

All and Weighted 모델은 source domain, target domain에 해당하는 모든 학습 데이터를 사용한다. 이 경우에는 보유하고 있는 모든 데이터를 사용한다는 측면에서는 긍정적이나 source domain 학습 데이터가 target domain 학습 데이터에 비해 너무 큰 경우에는 target domain의 특성이 반영되지 못하고 source domain 특성만 반영되어 실제로는 target domain 학습 데이터를 구축한 효과를 전혀 보지 못한다.

PRED 모델은 source 모델을 통해 구축된 분류기의 인식 결과를 target 분류기를 위한 학습 모델의 feature로 이용한다. 먼저 source 학습 데이터만으로 분류기를 학습한 후에, 이 분류기를 target data를 대상으로 실행한 결과 데이터를 얻는다. 그 다음 결과 데이터를 추가 feature로 하여 기존의 target data를 이용하여 학습을 하면 구축할 수 있다. PRED 모델 source, target data 구분하여 학습에 이용하는 장점이 있으나, 모델을 구축하기 위한 과정이 늘어나서 구축 속도가 느려지는 단점이 있다.

LININT 모델은 사용자가 파라미터 λ를 이용하여 source, target 데이터의 모델 반영 비율을 조정할 수 있어 데이터 상황에 맞게 유연하게 모델을 구축할 수 있지만, 이 역시 구축 속도가 느려지는 단점이 있다.

기계학습 기법이 효과적으로 동작하기 위해서는 충분한 학습 데이터가 필요하지만, 학습 데이터가 많아질 경우 학습 및 처리 시간이 늦어지는 것을 방지

하기 위해서 기계학습 알고리즘을 수정하기보다는 병렬처리 기반으로 데이터 구축이 필요하다.[빅 데이터 활용을 위한 기계학습 기술동향]

문서에 포함되어 있는 키워드를 기록할 뿐만 아니라, 문서 컬렉션 전체를 평가하여 어떤 문서가 비슷한 단어를 포함하고 있는지를 찾아낸다. LSI는 많은 단어를 공유하는 문서들이 의미적으로 가까운 것으로 간주하며, 공유하는 단어가 적으면 의미적으로 먼 것으로 여긴다. 단어들이 보여주는 패턴을 인식하여 유사성을 행렬 인수분해를 통해 밝혀내며, 이 방법은 앞에서 이야기한대로 행렬 인수분해를 통해 차원 축수를 하고 텀과 문서의 특징들을 축소된 차원에 표현하는 건데 안타깝게 계산량이 많은데에 반해 병렬처리가 어렵다고 한다. pLSI-조건부 확률로 계산이 가능한 방법이다. 두번째 방법은 병렬처리가 가능하여 많은 수의 사용자를 클러스터링하는 문제에서도 map-reduce 프레임웍을 통해 많은 시간을 단축시킬 수 있었다.
학습 데이터에 상당히 의존적인 알고리즘이다.

3.3 연구 개발의 필요성

폭증하는 데이터가 경제적 자산이 되는 빅 데이터의 활용에 기반 하였다. 기존과 차별화된 대용량 데이터의 새로운 분석과 추론을 통해 새로운 서비스를 개발할 수 있는 가능성이 무한하며 많은 양의 데이터는 새로운 정보를 발견할 가능성이 높다. 실생활 속에서 축적되는 다양한 유형의 데이터가 증가할수록 데이터의 활용가치는 정교해지고 세분화 되며 무한히 상승한다.[IT & Future STrategey 2012. 4 : 한국정보화진흥원]

위의 <표3-2>에서 볼 수 있듯이 빅 데이터의 현상은 이제 모든 분야에서 활발히 연구되고 상품화 하여 가치창출이 진행되고 있다. 국외를 보더라도 공공 및 기업 분야에서 빅 데이터 기술을 도입중인 반면에 국내 기업은 아직 미비한 단계이다. 디지털 정보의 양은 5년마다 10배씩 증가하고 있다. 소프트웨

어 프로그램들의 성능도 매우 좋아지고 있다. 지금까지 IT시스템의 역할은 정보를 조회하는 보조수단에 불과 했지만 빅 데이터를 기반으로 한 IT 시스템은 좀 더 세분화 되고 동적으로 현상 파악 및 대책을 제시한다. 빅 데이터 시대에 우리는 데이터 처리 활용 기술을 보다 발 빠르게 습득하여 대량의 데이터를 활용하여 가치를 창출하는 능력을 길러야 한다.

기업들은 소셜미디어에서 발생되는 각종 데이터들 속에서 소비자의 반응과 개인 프로파일이 결합된 패턴을 추출, 고객계층을 세분화시켜 타겟 마케팅을 실현할 수 있다. 빅 데이터는 수많은 데이터를 수집, 축적하는 것보다 무엇을 분석할 것인지를 분명하게 하여야 하며 통합적 사고와 해석능력이 중요하다. 수많은 데이터 속에서 목적에 부합하는 데이터를 찾아내고, 효과적인 분석과 분석결과를 제공하기 위해 최적화된 기계학습 알고리즘이 적용되는 빅 데이터 분석 플랫폼 구현이 필요하다. [빅 데이터가 여는 미래의 세상. 정보과학회지. 2012. 6]

빅 데이터 분석 · 예측을 위한 기계학습 및 인공지능 분야는 기반기술과 원천기술 개발이 진행되지 않아 기술 종속이 우려되는 분야다. 축적된 대규모 데이터 실시간 처리를 위해 하드웨어와 소프트웨어가 결합된 빅 데이터 고급 분석 기술과 추론과정의 지능화를 위한 기계학습 및 인공지능 기술 개발이 시급하다. 위에 언급된 도메인 적응 기법의 예와 같이 기계학습 그 자체를 개선하여 좀 더 효율적인 기법으로 빅 데이터 처리를 향상시키려는 노력이 필요하다.

4. 연구 목표

빅 데이터로부터 새로운 사회적 가치를 추출하기 위해서는 기존의 분석 방법이나 분석 툴이 아닌 새로운 접근 방식으로의 전환이 필요하다. 그러나, 우리나라의 경우 진정한 빅 데이터 활용사례는 매우 드물며, 국내 기업들의 빅 데이터 활용은 대부분 자체 분석기술이 아닌 해외 글로벌 기업의 기술을 빌리거나 오픈소스를 이용하고 있다. [빅 데이터 기술과 주요 이슈]

근래에 들어 기업들의 마케팅 전략은 소비자의 기능적 요구를 충분히 반영한 성능 및 가격 위주의 시장전략에서 제품 및 서비스에 대한 사용자의 편의성과 만족도를 극대화시키는 감성에 바탕을 둔 소비자 감성 지향형 전략으로 이동해 가고 있다. 감성은 소비자와 제품을 연결하는 새로운 패러다임이 되었으며 감성은 소비자를 사로잡는 아주 중요한 요소로 대두되고 있다. 이러한 시장의 요구에 발맞추어 본 연구는 빅 데이터 기반의 SNS 고객성향 분석을 통해 사용자의 감성적 측면을 반영한 여행지 선별 방법을 제안한다.

최근 텍스트 뿐 아니라 이미지, 멀티미디어, 소셜 미디어 등의 분석에서도 널리 쓰이고 있는, 대량의 텍스트 데이터에 내재되어 있는 시멘틱 패턴을 찾아내는 기법인 토픽모델링(Bayesian probabilistic topic modeling)을 고객 성향에 맞춘 여행지 선별을 위한 빅 데이터 분석 모델링에 사용한다. 토픽모델링 중 하나인 LDA(Latent Dirichlet allocation) 와 그의 추론 방법인 MCMC 샘플링 중심으로 토픽과 감성(sentiment)을 같이 다루는 기법을 제안하고자 한다.

LDA 알고리즘과 그것의 추론방법인 MCMC 샘플링 기법을 이용한 기계학습 알고리즘의 개선을 통해, 수집된 빅 데이터로부터 고객성향에 따른 여행지 선별을 위해 소비자 구매, 취향, 감정, 정서 패턴을 효과적으로 파악하고 분석하는 감성과 토픽이 결합된 최적화된 ML알고리즘을 구현하고, 빅 데이터 분석

플랫폼에 적용하여 소비자 취향에 맞는 다양한 요구 분석이 가능하도록 하는 개인화 서비스까지 제공할 수 있도록 한다.

기계학습 중 감독학습(supervised learning)의 경우 감독학습을 위한 학습 데이터를 구축하는 것이 시간과 비용이 많이 들고, 이렇게 구축된 학습 데이터도 적용 분야(domain)가 변경될 경우 다시 재구축 비용이 발생하는 것이 가장 큰 문제이다. 학습 데이터 적용 분야가 변경될 때마다 각각의 분류기를 구축해야하는데, 이것은 현실적으로 불가능하다. 이런 문제를 도메인 적응 문제라고 정의하는데, 이를 해결하기 위해 위에서 언급한 무감독 학습법(Unsupervised learning)인 LDA 모델 기법을 적용하여 다른 도메인에도 쉽게 적응할 수 있도록 한다.

국내 전문가들은, 한국 시장의 경우 소수의 대형 기업이 시장을 분할해 독식하는 구조 때문에 경쟁우위 확보에 대한 욕구가 상대적으로 부족하므로 일반 기업의 빅 데이터 도입 규모는 크지 않을 것으로 전망한다. 하지만 대규모의 데이터웨어하우스 시스템 투자에 대한 기업의 부담을 줄이고, 하둡(Hadoop)이나 NoSQL 같은 SW자체의 가격이 상대적으로 저렴하고, 규모 확장이 쉬운 빅 데이터 분석 플랫폼을 구현하여 작은 규모의 구축 시스템으로도 SNS 고객 성향에 따른 여행지 선별과 같은 비즈니스 모델의 빅 데이터 분석이 가능할 수 있다는 것을 보이고자 한다.

또한, 빅 데이터 분석 플랫폼을 향후 다양한 분야에 적용하여 효율적인 빅 데이터의 분석과 비즈니스 활용 모델로서 기업 또는 국민 경제의 경쟁력강화 및 생산성 향상 극대화를 도모하고자 한다.

5. 연구 내용

빅 데이터 시대가 도래하면서 보다 많은 정보들을 사용할 수 있게 되었다. 이 많은 정보들 중에서 우리가 원하는 정보를 찾아낸다는 건 더욱 더 어려워졌다. 대량의 정보를 정리하고, 검색하고, 이해할 수 있게 도와주는 도구가 필수적으로 필요하다. 이 역할을 해 줄 도구가 바로 '토픽 모델링'이다. 토픽 모델링은 많은 자료들 중에서 특정 주제(Theme)를 찾아내서 찾아낸 주제를 바탕으로 문서들에 특정 주석(Annotate)을 달아준다. 문서에 달린 주석을 바탕으로 문서를 예측하여 정리하고 요약 과정으로 진행된다. 최근 가장 주목을 받고 있는 토픽 모델링은 바로 LDA(Latent Dirichlet Allocation)이다.

5.1 연구 개발내용

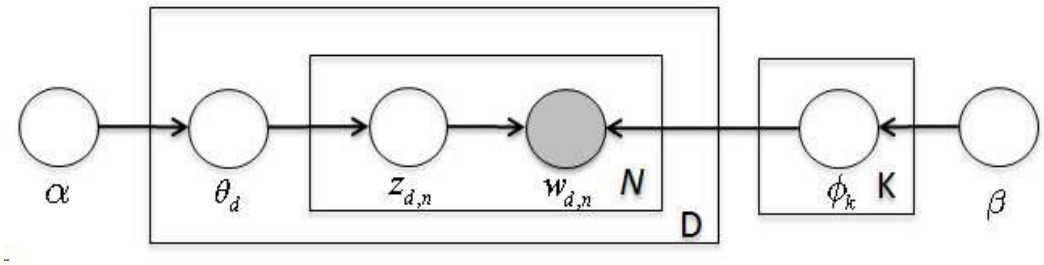
사용자의 감성적 측면을 반영한 여행지 선별을 위해 LDA 토픽모델링과 그의 추론 방법인 MCMC 샘플링 기법 중심으로 감성(sentiment) 결합 토픽 모델링을 활용한다.

가. LDA 기법

자연어 처리에서 잠재 디리클레 할당(Latent Dirichlet allocation, LDA)은 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지에 대한 확률 모형이다. 미리 알고 있는 주제별 단어수 분포를 바탕으로, 주어진 문서에서 발견된 단어수 분포를 분석함으로서 해당 문서가 어떤 주제들을 함께 다루고 있을지를 예측할 수 있다. LDA에는 몇 가지 가정이 있는데 그 중 중요한 것은 단어의 교환성(exchangeability)이다. 이는 '단어 주머니(bag of words)'라고 표현하기도 하는데 교환성은 단어들의 순서는 상관하지 않고 오로지 단어들의 유무만이 중요하다는 것이다. LDA는 단순히 문서의 주제를 찾는데 쓰이는 것

이 아니라 이미지, 소리 등 다양한 영역에서 쓰일 수 있다.[위키백과]

문서 내 단어 정보를 바탕으로 문서의 은닉 변수에 해당하는 토픽을 밝혀내는 무감독 학습법으로 주목을 받고 있는 기법이다. 여기서 은닉 변수인 토픽은 의미적으로 유사한 문서에서 자주 함께 등장하는 단어들의 집합이며, 수작업으로 문서에 대한 정보를 추가하지 않은 오직 단어 사이의 동시 발생 빈도를 바탕으로 은닉 변수를 학습한다. LDA는 문서의 은닉 토픽을 찾기 위한 기존 방법인 Latent Semantic Indexing(LSI), 확률적 LSI(pLSI) 방법의 약점인 과적합(overfitting) 현상 및 데이터 증가에 따른 모델 매개변수 증가 현상을 해결하면서도 다층 베이지안 모델(Hierarchical Bayesian models)을 기계 학습 분야에 적용한 좋은 예이다.



<그림5-1> LDA 모델

<그림5-1>는 LDA 모델의 그래프 표현이다. 전체 데이터는 D개의 문서로 이루어져 있으며, 각 문서 d는 N_d개의 단어 $w = w_1, w_2, \cdots, w_{N_d}$로 구성된다. 하나의 문서는 우리가 관측할 수 없는 다수의 토픽z로 구성되어 있으며, 문서의 토픽 분포에 따라 생성되는 단어의 빈도가 결정되는 것이라고 할 수 있다.

1. Choose N ~ Poisson(ξ).
2. Choose Θ ~ Dir(α).
3. For each of the N words w_n :
   (a) Choose a topic z_n ~ Multinomial(Θ).
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial

probability conditioned on the topic z_n

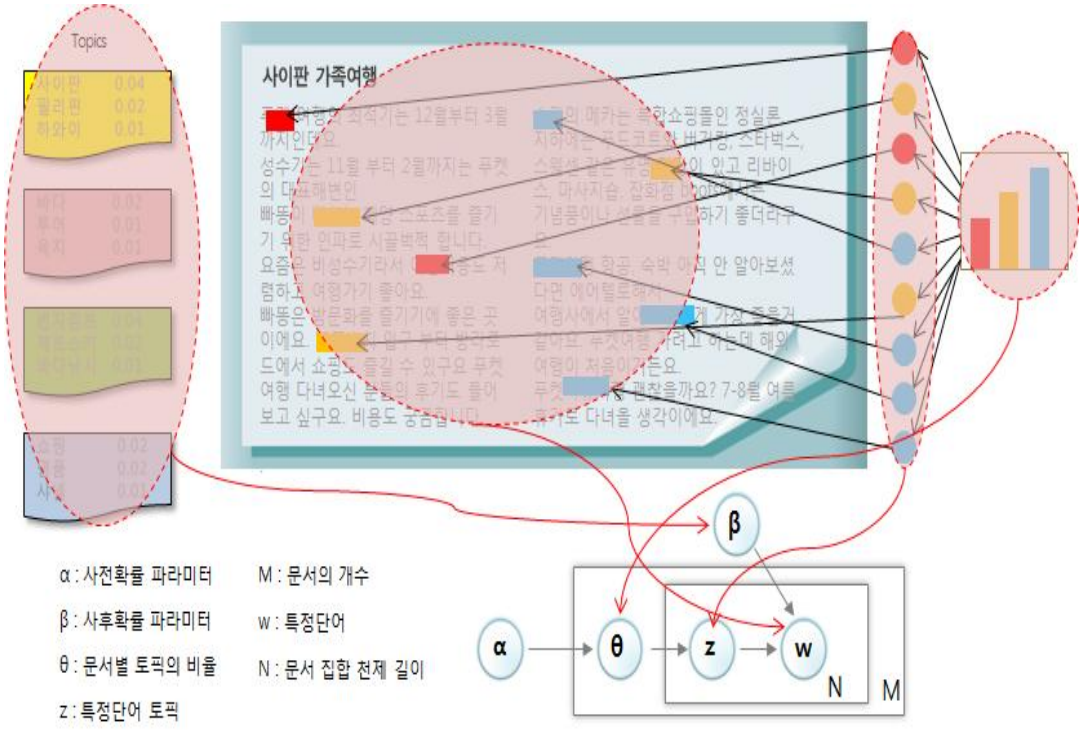
α와 β는 코퍼스 단위로 정해지는 값이고, N과 Θ는 문서 단위로 정해지는 값이다. β는 각 주제별로 특정 단어가 생성될 확률이 담긴 테이블이며, N은 문서의 길이, Θ는 해당 문서에서 각 주제의 비율를 나타낸다. zi는 문서의 i번째 단어에 대한 주제 벡터(하나의 엔트리만 1이고 나머지는 0)이다. 이 모델에서 주제의 개수는 K로 고정되어 있으며, 따라서 Θ와 zi는 길이가 K인 벡터이다.

전체 문서는 K개 토픽으로 표현 가능하다고 할 때, 먼저 하나의 문서에 포함된 토픽의 분포 θ를 디리슐레(Dirichlet)분포로 정의한다. 그 뒤 문서 내의 모든 단어에 대해 a,b 과정을 반복한다. 먼저 토픽 분포 θ를 매개변수로 하는 다항 분포로부터 토픽 인덱스를 선택한다.

보유한 많은 텍스트 기초에 α와 β를 알아 두고, 개별 문서의 θ를 계산할 수 있으면, 이 θ를 갖고 유사도 계산이나 분류 작업을 훨씬 쉽고 정확하게 해낼 수 있다.

$$p(z_1,\cdots,z_n)=\int p(\theta)(\prod_{n=1}^{N}p(z_n|\theta)$$
$$p(w,z)=\int p(\theta)(\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n))d\theta$$

첫 번째 식은 문서의 토픽 생성, 두 번째 식은 문서의 토픽과 단어 생성을 나타낸다. 문서 토픽을 나타내는 z는 θ에 대한 조건부 확률이다.



<그림5-2> LDA 모델의 생성과정

<그림5-2>는 LDA 모델의 생성과정이다. 단어들은 특정 토픽들 z로부터 생성되고, 해당 문서가 어떤 토픽 비율 θ를 가질 것인지 파라미터가 α인 Dirichlet distribution에 의해 결정된다. 토픽 z는 토픽 자체라기보다 토픽을 나타내는 숫자로 볼수 있다. 문서 내의 각 단어 별로 θ의 비율로 토픽 z을 선택하게 되는 것이다.[토픽모델링을 사용한 『독립신문』 논설의 분석 및 저자 판별도구 2014. 3 : 김기남, 박호식]

나. 추론방법 MCMC 샘플링 기법-깁스 샘플러(Gibbs sampler)

베이지안 확률 모델에서 관측 데이터가 주어졌을 때 비관측 데이터에 대한 확률을 계산하는 과정을 추론(inference)라고 한다. 앞서 설명한 모델은 단어들이 주어졌을 때 추론을 통해 문서의 토픽 분포, 각 단어가 속한 토픽 및 각각 토픽의 단어 분포에 대한 사후 분포를 계산한다. 사후 분포를 계산하기 위해서 사용되는 것이 바로 베이즈 법칙(Bayes' rule)이다. H는 모델이고, D는

데이터를 나타내는 확률변수(random variable)라고 할 때, 베이즈 법칙은 다음과 같다.

$$p(H|D) = \frac{p(H,D)}{\sum_{h \in H} p(H,D)}$$

실제적으로 다뤄지는 대부분의 확률 모델은 정확 추론(exact inference)이 불가능하다. 그렇기 때문에 원래 모델의 근사형태로 접근해야한다. 이 때, 사용할 수 있는 방법 중 하나는 결정적(deterministic) 근사 기반의 추론이며, 다른 하나는 샘플링 기반의 추론이다.

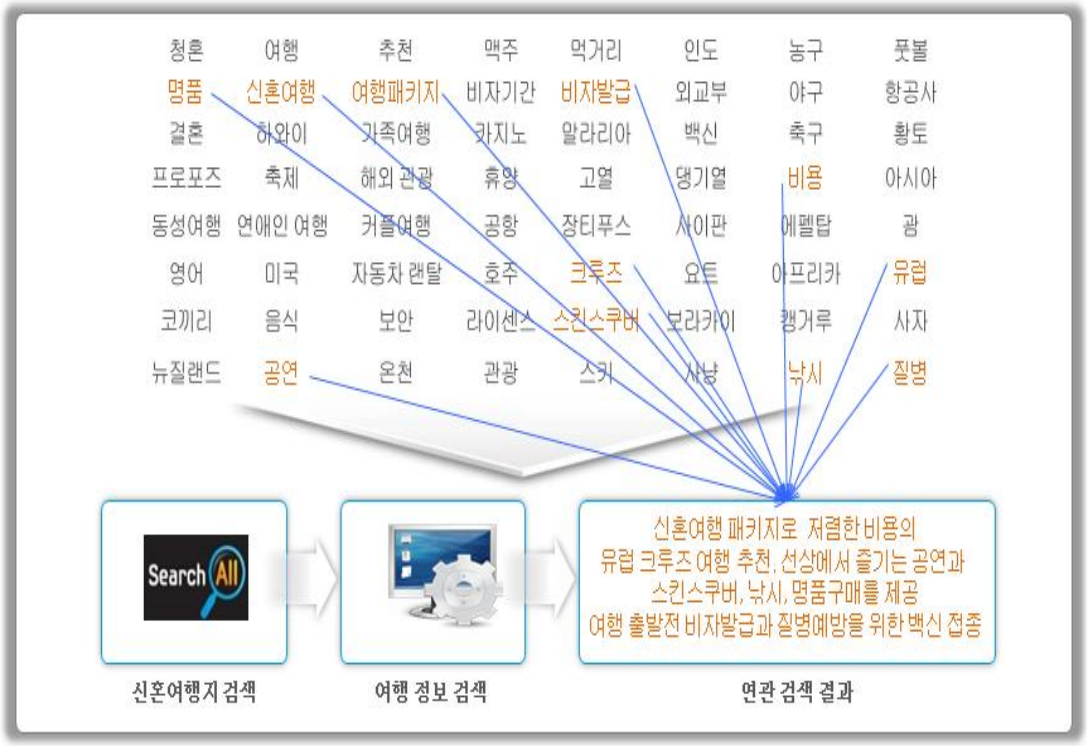
먼저 모든 은닉 변수 θ, ϕ, z에 대한 사후 분포를 결정하기 위해서는 많은 계산을 필요로 하기 때문에, 가장 관심 있는 확률 변수인 토픽 z에 대한 사후 분포로 결정한다. 다른 은닉 변수인 θ, ϕ에 대해 각각 적분한 결과로 얻을 수 있는 가장 관심 있는 은닉 토픽 z에 대한 사후 분포를 다음과 같이 정의할 수 있다.

$$p(\boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\boldsymbol{w}|\boldsymbol{z}, \beta)p(\boldsymbol{z}|\alpha)}{\sum_{z} p(\boldsymbol{w}|\boldsymbol{z}, \beta)p(\boldsymbol{z}|\alpha)}$$

깁스 샘플러는 특히 비관측 확률변수가 여러 개인 경우에 다른 샘플링 방법보다 효과적으로 사용된다. 하나의 문서가 N 개의 단어로 구성되어 있을 때, N 개의 은닉 토픽에 대한 사후 분포를 다음과 같이 계산한다.

$$\begin{aligned} z_1 &\sim p(z_1 | \boldsymbol{z_{-1}}, \boldsymbol{w}) \\ z_2 &\sim p(z_2 | \boldsymbol{z_{-2}}, \boldsymbol{w}) \\ &\vdots \\ z_N &\sim p(z_N | \boldsymbol{z_{-N}}, \boldsymbol{w}) \end{aligned}$$

전체 N 개 단어의 토픽을 결정하는 과정을 샘플 추출 과정이라고 할 때, 미리 지정된 반복횟수만큼 샘플 추출 과정의 반복을 통해 사후 분포 계산이 가능하다. 위에서 정의한 사후 분포를 계산을 통해 알 수 있다면, 우리는 단어가 주어졌을 때 어울리는 토픽이 무엇인지 확인할 수 있다.



<그림5-3> LDA알고리즘을 이용하여 여행지 검색 예시

<그림5-3>의 윗 부분은 주제별로 관련성이 높은 단어를 뽑은 것, 즉 β이고, 아랫부분은 예제 문서에서 주제가 확실한 단어에 색칠한 것, 즉 ϕ를 보여준다. 윗부분에서 토픽 이름은 알고리즘이 자동으로 뽑아주는 것이 아니라 사람이 정한 것이다. 비슷한 단어들이 엮여 나오기 때문에 문서 모델링 뿐 아니라 키워드 클러스터링에도 LDA를 이용할 수 있다.

사실 LDA는 문서를 다룰 때만 쓰이는 것이 아니라, 비슷한 형태의 모든 데이터 분석에 활용할 수 있다.

<그림5-4> 빅 데이터 분석 플랫폼

① 빅 데이터 수집 통합 기술-데이터의 형태과 소재에 무관하게 데이터를 확보하는 기술

- 크롤링-검색엔진 로봇을 이용한 새로운 데이터 생성 : 수집대상 식별
- 소스데이터 추출. 이동. 변환. 적재
- 내·외부 이종 데이터 통합
- 데이터 가상화
- NoSQL(HBase)
- 빅 데이터 저장-분산파일 시스템 이용(HDFS-Hadoop Distributed File System)

② 빅 데이터의 전처리 기술-지속적으로 발생하는 비정형 스트림 데이터를 정제하고 구조화하여, 분석의 정확성을 높이고, 심층 분석을 가능하게 하는 기술

- 비정형 데이터 처리-기계학습이나 텍스트 마이닝을 통한 반정형화/정형화 변환
- 필터링(데이터의 오류발견, 보정, 삭제 및 중복 데이터의 삭제)을 통한 데이터품질 향상
- 관리 및 모니터링을 통한 연산/처리
- 데이터 통합 및 익명화
- 데이터 정제

③ 데이터 저장/관리 기술-웹 데이터, 소셜 미디어, 센싱 정보 등의 폭증하는 다양한 형식의 데이터를 실시간 저장/관리할 수 있는 분산 컴퓨팅 기술

- 대용량 분산 파일 시스템(MapReduce/Hive/Pig)
- 데이터 분산. 병렬 처리
- NoSQL
- 인-메모리 DB
- 인-DB 분석
- Indexing/Searching

④ 데이터 분석 기술-빅 데이터에 내재된 가치를 추출하기 위해 필요한 대규모 통계처리, 데이터 마이닝, 그래프 마이닝 등의 분석 방법, 기계학습 및 인공지능을 활용한 심층 분석 기술

- Descriptive Analysis
- Predictive Analysis
- Knowledge Base(DSS)
- **LDA기법을 적용한 감성결합 토픽모델링**
- 자연어 처리
- 텍스트 마이닝
- Contents Analysis

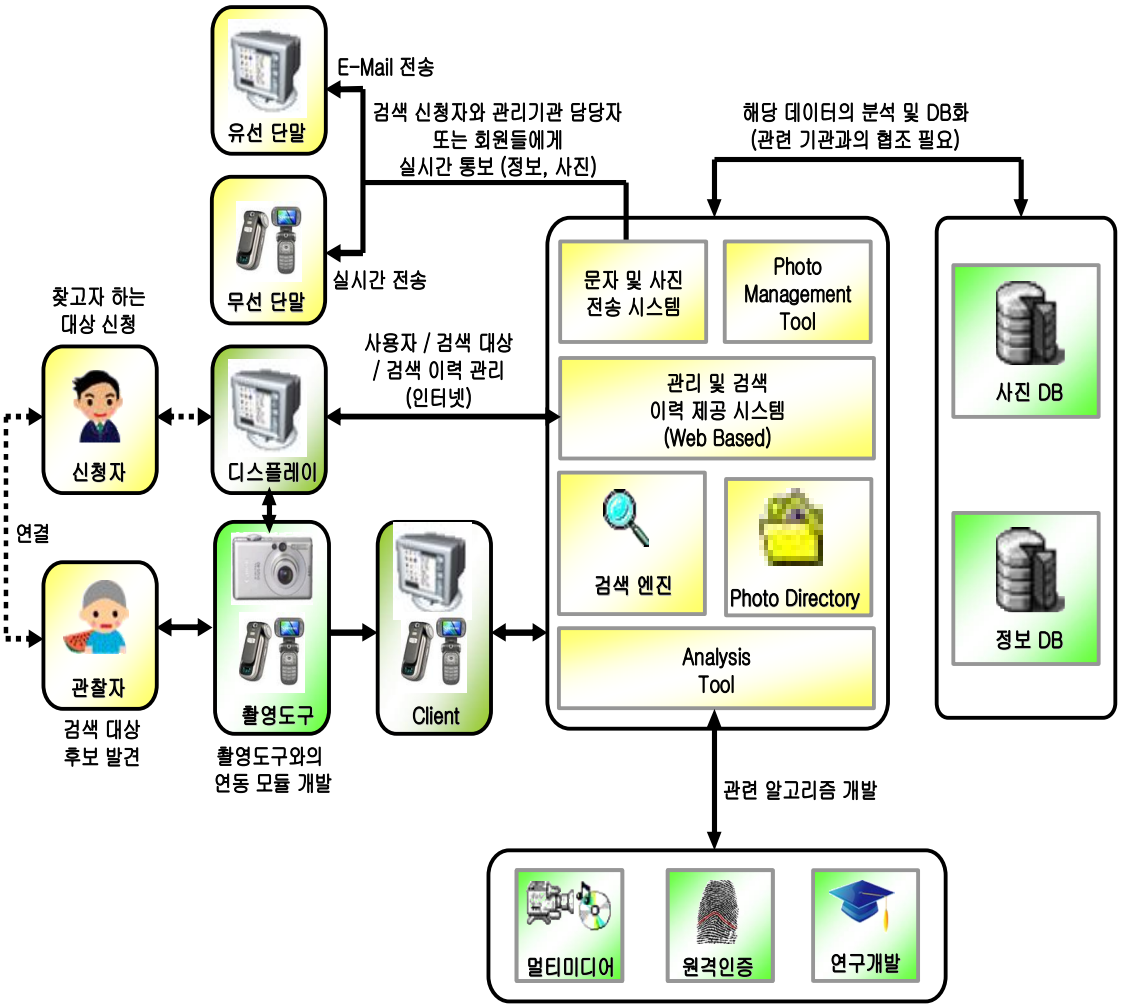
・Mahout(다양한 Machine Learning 알고리즘을 라이브러리 형태로 제공-대용량의 데이터를 필요로 하는 지능형 애플리케이션 개발을 위한 분산/병렬처리가 가능한 기계학습 라이브러리)

⑤ 데이터 분석 가시화 기술-비전문가가 데이터 분석을 수행할 수 있는 환경을 제공하는 분석 도구 기술과 분석 결과를 함축적으로 표시하고, 직관적인 정보를 제공하는 인포그래픽스 기술로 구성된다.

・데이터를 실세계처럼 구현
・결과의 시각화
・분석 자연어 처리
・그래픽 기반 모델링 도구
・분석알고리즘 자동 실행도구
・인포그래픽스
・실시간 가시화 도구
・동적 가시화 도구

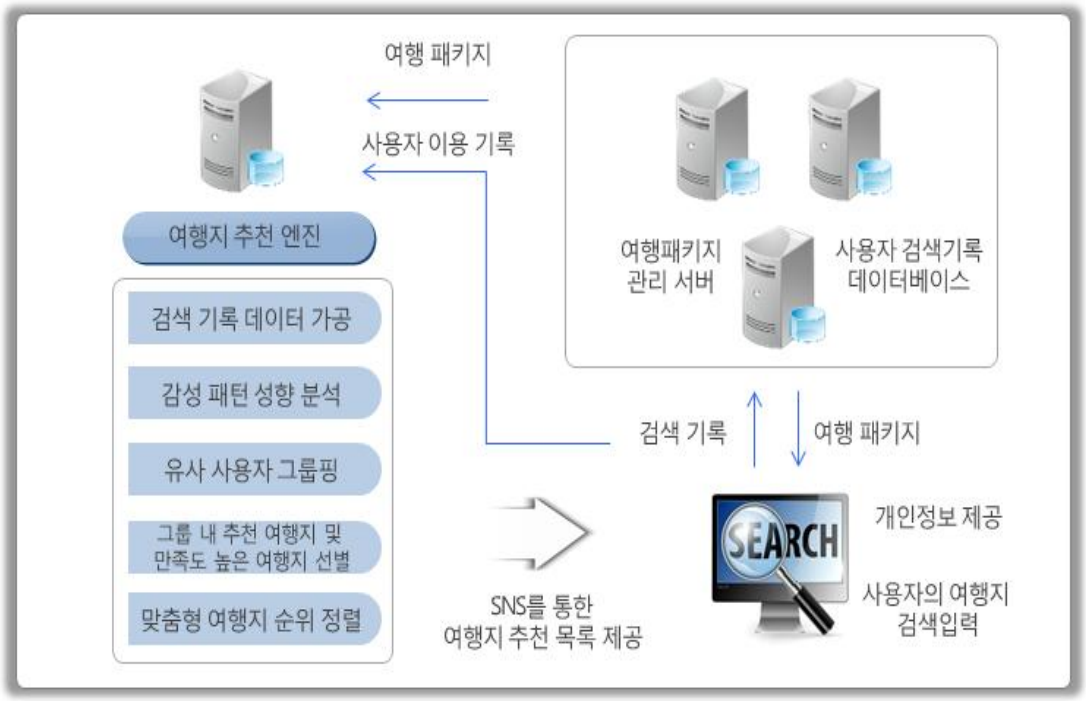
5.2 추진전략 및 방법

빅 데이터 분석 기반의 SNS 고객여행지 선정 모델에 대한 추진전략 및 방법은 <그림5-5>와 같다



<그림5-5> 모바일 객체인증 시스템 구성 및 활용 흐름도

여행지 검색 요청이 접수 되었을 때 여행지 패키지를 자동 추천하여 제공하는 여행지 자동 추천시스템의 예상 구성도는 아래 <그림5-6>과 같다.[빅데이터 분석 기반의 SMS 고객선정 프로파일링 모델에 대한 실증적 연구 2012. 6 : 김종현]



<그림5-6> 여행지 추천 가상 시스템 구성도

6. 연구 결과

6.1 연구결과의 활용방안

가. 학문적 측면

○ 다양한 매체로부터 수집된 데이터 간 연관성을 통한 의미 있는 정보를 분류할 수 있는 기계학습 알고리즘의 개선

○ 전문적인 연구를 통한 난제 해결의 방법론 제시

- 단일 학문 분야로는 풀 수 없는 문제의 해결을 위해 두 개 이상의 학문으로부터 나오는 정보, 데이터, 기술, 도구, 관점, 개념, 이론 등을 유기적으로 결합하여 연구 수행
- BT 외에 NT, ET, CT 뿐만 아니라 금융, 건설 유통, 자동차, 항공, 로봇 분야의 융·복합 학문에 기여

나. 기술적 측면

○ 빅 데이터 처리에 사용되는 기계학습 기법 그 자체를 개선하여 좀 더 효과적인 동작 가능

○ 빅 데이터 분석을 통한 효율화. 개인화. 선제적인 비즈니스 모델 혁신

- 효율화 : 기존 경영 의사결정은 오랜 경험, 최고 경영자의 직관, 과거 데이터를 통한 트렌드 분석에 의존하였으나, 빅 데이터를 기반으로 과거 및 현재 현상을 파악하고 물류, 재무, 기획, 마케팅 등 경영 전반의 데이터가 실시간으로 분석되고 이를 통해 최선의 의사결정을 제공할 수 있을 것으로 예상된다.

- 개인화 : 온라인 상에서의 이용자 활동 정보, SNS 등을 통해 축적된 개인

정보 등이 결합되어 사용자 개인에 특화된 서비스를 제공할 수 있을 것으로 기대되고, 현재 활용되고 있는 광고 분야 뿐 아니라 의료, 교육 등 모든 서비스 분야로 확대가 가능할 것으로 예상된다.

- 미래예측력 : 과거 및 실시간 데이터 분석을 통한 축적된 개인정보를 바탕으로 개인 또는 조직 전체의 행동 및 의사결정 패턴을 도출할 수 있게 되고, 이러한 분석을 통해 미래 적용 가능한 시나리오를 제시하여 예측 가능한 행동 및 발생가능한 문제점을 사전에 방지하는 서비스가 가능하게 될 것이다.

6.2 기대성과

○ 빅 데이터는 새로운 혁신 동력이자 새로운 비즈니스를 위한 플랫폼, 그리고 기업과 정부의 성장 동력으로 활용이 가능하다. 빅 데이터의 핵심은 빅 데이터 자체에서부터 이를 활용한 사용자 어플리케이션에까지 광범위하며 특정 서비스, 물리적 하드웨어가 아닌 빅 데이터 에코시스템과 플랫폼으로 확장될 수 있다.

○ 빅 데이터 분석 자체가 새로운 비즈니스로 각광

- 축적된 빅 데이터를 유통하거나, 빅 데이터를 가공해 시각화 및 수치화된 2차 데이터를 판매하는 새로운 비즈니스가 등장하게 될 것이고, 빅 데이터를 통한 개인화, 효율화 등의 기존 비즈니스 혁신을 유도하는 빅 데이터 컨설팅이 주목받게 될 것이다.

○ 통상의 빅 데이터를 수집하고 분석하는 절차에 플랫폼 구현을 위한 학문·기술적 장치를 추가해야 하는 연구 범위를 갖기에 연구 과정에 투입된 연구 인력은 빅 데이터 처리 및 분석영역에서 현장 맞춤형 전문 인력으로 성장할 것으로 기대된다

7. 결론

LDA 모델링 기법을 활요하여 고객성향을 바탕으로 최적의 여행지 조건을 제시 할 수 있음을 제안하였다.

LDA 모델링 기법이 적용되고 있는 사례가 지속적으로 나오고 있으며, BigData 부분은 새로운 창조보다는 수많은 정보들로부터 누가 먼저, 무엇을, 어떤 정보를, 어떻게 가공하여 어떤 비즈니스를 선점 하느냐가 최고의 이슈일 것이다.

수많은 사람들이 경험한 데이터를 토대로 이미 검증된 최고 수준의 정보제공은 기업 성장에 발판이 되고 고객의 편의성을 극대화 할 수 있으며 수많은 시행착오를 최소화 함으로써 불필요한 비용과 인력낭비를 방지 할 수 있다.

이종간 정보 결합(자동차, 의료, 환경, 신기술, IT등)을 통하여 새로운 가치 창출할 수 있다.

자동차를 운전중인 운전자의 운전패턴, 도로상황, 지역특색, 자동차 상태 등의 정보를 수집하고, 바이오리듬을 활용한 운전자의 건강상태, 혈압, 맥박수, 배고픔, 스트레스 지수 정보를 수집하고, 외부 온도, 습도, 날씨 바람세기 등을 수집하여, 수집된 정보를 분석하고 지속적은 사용자 모니터링을 통하여 사용자가 현재 무엇을 필요로 하는지 상황에 맞는 대안을 즉시 제시 하는 등 신개념 융합시장도 창출할 수 있다

Travel SNS recommend using the LDA-based data modeling techniques

Kabsu Jung

Department of Electric and Electronics and Computer Engineering
Graduate School of Industry and Technology
Chonnam National University
(Supervised by Professor Ji-seung Nam)

(Abstract)

In recent years, the tourism industry is the proportion who purchase select the destination and travel products through the Tourist Information Retrieval Using SNS as popular IT technology and information retrieval using the Internet is going higher. Moreover, due to the widespread smartphone anywhere, anytime information and content, such as providing travel information and travel experiences later by the nature of SNS, which can be easily obtained.

Real-time nature of consumer sentiment, identify the emotional patterns Twitter as a way to provide a new destination of the individual tendencies custom screening services, blog, cafes, it is a very efficient use of Big Data on SNS. However, the big amount of data that is only significant information as well as real-time generation of data to be updated, and a wide variety of atypical and velocity, the unstructured

data Because of the complex nature that are mutually fused to the existing data analysis techniques are not sufficient.

Latest images as well as text, multimedia, social media, widely used in analysis such as Big fit for travel screening techniques to find patterns that are inherent in a large number of semantic text data, topic modeling (Bayesian probabilistic topic modeling) a tendency to customers the modeling used for data analysis. One of the topics of modeling LDA (Latent Dirichlet allocation) and his way of reasoning centered on MCMC sampling techniques to deal with such an emotional topic (sentiment) is proposed.

In this paper, we want to promote competitiveness and productivity maximization of the company or the national economy as a model for efficient analysis and business leverage big data applied to various fields next to big data analytics platform.

참고문헌

[1] 정용찬, “빅 데이터 혁명과 미디어 정책 이슈” (KISDI Premium Report 12-02). 정보통신정책연구원. pp.4

[2] 김성태,“빅 데이터 시대! SNS의 진화와 공공정책”,한국정보화진흥원, 제13호, pp.1 ,2012.11.

[3] 류한석, “데이터비즈니스의 이슈와 전망”."DIGIECO", https://www.digieco.co.kr

[4] 김한나, “ 빅 데이터의 동향 및 시사점”, 제24권 19호 통권 541호, pp. 51. 2012. 국가정보화 백서 재구성, pp.53-56. 2012.

[5] 박준규, “빅 데이터를 위한 분석기술 활용방안 연구”,세종대학교, 석사학위논문, pp.34, 2012.

[6] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation, Journal of Machine Learning Research 3", pp.993 - 1022, 2003.

[7] 김성태, “성공적인 빅 데이터 활용을 위한 3대 요소 : 자원, 기술, 인력”한국정보화진흥원, 제3호, pp.2, 2012.

[8] 김상락, 강만모, 박상무, “빅 데이터가 여는 미래의 세상”, 정보과학회지.pp.20 2012.

[9] 정지선, 新가치창출 엔진, 빅데이터의 새로운 가능성과 대응 전략, 2011.

[10] 김기남,박호식, "토픽모델링을 사용한 『독립신문』 논설의 분석 및 저자 판별도구" pp.9, 2014.3.

[11] 김종현, “빅데이터 분석 기반의 SMS 고객선정 프로파일링 모델에 대한 실증적 연구",숭실대학교, 박사학위논문, pp.7-14. 2012. 6