빅데이터 분석 기반의 제품 평판 마이닝 알고리즘

박상민*, 박새빛**, 온병원***[1)]
군산대학교 통계컴퓨터과학과
e-mail : bk1162@kunsan.ac.kr*, 1400662@kunsan.ac.kr**,
bwon@kunsan.ac.kr***

An algorithm for mining the reputation of a product based on big data analytics

Sang-Min Park*, Sae-Bit Park**, Byung-Won On***
Department of Statistics and Computer Science, Kunsan National University

요 약

최근 여론조사 분야에서 빅데이터 분석 기법이 널리 활용되고 있다. 기업에서는 최근 출시된 제품에 대한 선호도를 조사하기 위해 기존의 설문조사나 전문가의 의견을 단순 취합하는 것이 아니라, 온라인상에 존재하는 다양한 종류의 데이터를 수집하고 분석하여 제품에 대한 대중의 기호를 정확히 파악할 수 있는 방안이 필요하다. 본 연구에서는 빅데이터로부터 제품의 평판을 자동으로 찾아내는 텍스트 마이닝 방안을 제안하고, 소나타 자동차를 중심으로 제안 방안의 효율성을 평가하고 실험 결과를 자세히 분석한다.

1. 서론

2016년 아시안리더십콘퍼런스에 참여한 짐 클리프턴 갤럽 회장은 “모바일 기기의 확산으로 인해 갈수록 사람들은 여론조사에 덜 협조적이어서 여론조사로 앞일을 예측하는 것이 과거 그 어느 때보다 어렵다”라고 밝혔다. 그러면서, “설문조사를 통해 얻는 데이터 자체는 이제 큰 쓸모가 없고 신뢰하기도 어렵다. 빅데이터 분석을 통해 의미를 파악하고 새로운 발견이나 해법을 제공하는 것이 미래의 갤럽이 할 일”이라고 언급했다[8]. 이와 같이 조사원이 여론조사를 직접 수행하고 결과를 분석하는 기존의 전통적인 방식은 효율성이 크게 떨어지고 있다. 조사원과 통계 전문가를 활용하기 위해서는 많은 비용이 들어가고, 같은 내용이라도 설문 문항의 차이로 인해 다른 결과를 얻거나 설문 작성자의 주관적인 판단이 들어갈 위험성이 있다. 무엇보다 중요한 것은 표본의 크기가 작고 응답률이 높지 않으면 모집단을 추정하는데 많은 왜곡이 가해질 수 있어 궁극적으로 결과에 대한 신뢰를 얻을 수 없다. 게다가 여론조사를 신속하게 진행하는 것은 앞날을 예측하는데 매우 중요하지만, 사람에 의한 조사는 빠른 시간 내에 결과를 얻기는 쉽지 않는 것이 현 실정이다. 이러한 여러 제약을 해결하기 위해서 최근에는 빅데이터 분석 기법을 도입하여 인터넷상에 있는 정보를 수집하고 통계 분석을 통해 여론조사 결과를 도출하려는 흐름이 일고 있지만, 구체적인 방법론에 대한 심도 있는 연구가 학계 차원에서 이루어진 적은 없다.

본 논문에서는 다양한 여론조사 분야 중에서, 특정 제품에 대한 대중의 선호도를 파악하는 텍스트 마이닝 방안을 제안한다. 이러한 문제를 ‘제품 평판 마이닝’(Mining the reputation of a product)이라고 명명하고, 구체적인 방법론과 특정 제품에 대한 사례 연구를 통하여 제안 방안의 효용성을 입증하고, 실험 결과로부터 의미 있는 지식을 추출하고자 한다. 먼저 사례 연구를 위해 소나타 자동차를 분석 대상으로 선정하고, 본 논문에서 제안하는 방안을 적용하여 얻어진 결과를 분석하였다. 소나타 자동차는 그 어떤 제품보다 대중의 관심이 많은 아이템이며, 지난 수십 년 동안 꾸준하게 팔린 베스트셀러이자 스테디셀러이다. 따라서 대중의 관심이 많고 인터넷상에 관련 정보도 쉽게 얻을 수 있다. 특히 대중의 관심이 바로 투영되어 나타나는 곳은 사용자 후기 게시판이다. 자동차를 구입하거나 관심이 많은 고객들은 후기 사이트에서 정보를 얻거나 제품의 단점을 지적하여 많은 사람의 여론을 환기시킨다. 본 연구에서는 국내 최대 자동차 후기 게시판으로 널리 알려진 ‘보배드림’ 웹 사이트로부터 사용자 후기 게시판에 실린 텍스트 데이터를 사용하였다[5]. 제안 알고리즘을 보배드림과 같은 문서 코퍼스(document corpus)로부터 주요 토픽을 추출한다. 토픽의 예로서 자동차의 품질, 가격, 디자인 등을 둘 수 있다. 각 토픽에 대해 구체적으로 어떤 이야기들이 화자 되는지를 요약해서 보여주는 알고리즘이 필요하다. 보배드림에서는 소나타의 디자인이라는 토픽에서 구체적으로 어떤 이야기들이 사람들에게 많이 언급되는지를 파악하는 것이다. 또한 토픽의 긍부정을 판단하여 대중의 제품에 대한 선호도를 토픽별로 쉽게 알 수 있으며, 제품을 만든 회사 경영진에게는 중요한 정보가 될 것이다. 예를 들어, 소나타 자동차의 디자인 토픽에 대한 긍부정의 통계 수치를 계산하고, 어떤 점이 긍정이고 부정인지를 파악할 수 있다면, 제품을 개선하거나 제품 홍보를 하는데 크게 도움이 될 것이

1) 군산대학교 통계컴퓨터과학과 조교수, 교신저자(Corresponding Author)

다.

본 논문의 구성은 다음과 같다. 2장 제안 방안에서는 제품의 평판 마이닝 알고리즘에 대해 구체적으로 설명한다. 3장에서는 실험 환경 및 결과에 대해 자세히 논의한다. 4장에서는 본 연구의 기초가 되거나 관련 있는 연구를 정리하여 소개하고, 5장에서 결론 및 향후 연구 방향에 대해 기술한다.

2. 제안방안

그림 1은 제안 방안의 개요를 나타낸다. 제안 알고리즘은 (1) 토픽 추출 (2) 토픽 요약 (3) 토픽 선호 판별 등으로 구성된다. OutWit Hub[4]라는 소프트웨어를 사용하여 보배드림에서 총 2,532개의 사용자 후기 게시판의 문서를 자동으로 수집하였다. 이러한 사용자 후기 텍스트 문서들을 입력으로 받아, 먼저 주요 토픽을 추출한다. 주요 토픽의 예로는 소나타 자동차의 품질, 서비스, 또는 디자인 등이 될 것이다. 각 토픽에 대해서 일반 사용자들이 주로 이야기하고 있는 내용을 요약해서 보여주는 토픽 요약 과정이 수행된다. 끝으로 각 토픽의 선호를 측정하고, 긍정적인 내용은 무엇이 있는지, 부정적인 내용은 어떠한 것이 있는지를 요약해서 보여준다. 각 구성요소의 구체적인 알고리즘은 다음 장에서 설명한다.

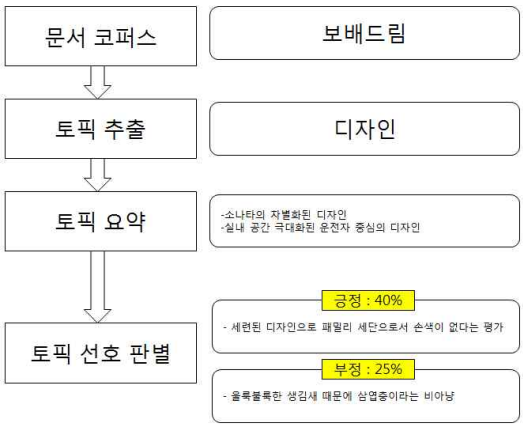

그림 1 제품 평판 마이닝 알고리즘

2.1 토픽 추출

토픽 추출을 위해 본 논문에서 잠재디리클레할당(Latent Dirichlet Allocation; LDA) 방법을 사용한다. LDA는 텍스트 문서들의 주요 토픽을 추출하는 통계적인 기법 중에 하나이다[2]. 이 방법은 하나의 텍스트 문서는 여러 토픽들이 혼합되어 있다는 가정에서 출발하며, 어떤 텍스트 문서도 생성되기 전에 이미 토픽 구조가 존재하며, 숨겨져 있는 토픽 구조(hidden topic structure)에 의해서 문서가 생성된다는 생성 확률모델(generative model) 기반으로 동작한다. 숨겨져 있는 토픽 구조로는 토픽의 개수, 문서 d의 토픽 분포($\theta^{(d)}$), 문서 d에 있는 각 단어의 특정 토픽 배정 확률($Z^{(d)}$), 토픽 z의 단어 분포 ($\Phi^{(z)}$) 등이 있다. α와 β의 사전 확률이 주어지면, 문서 d의 사후 확률은 다음과 같이 계산된다.

$$P(d,w) = P(d) * P(\theta^{(d)}|\alpha) * \sum_{z} P(\Phi^{(z)}|\beta) * P(w|z,\Phi^{(z)}) * P(z|\theta^{(d)}) \quad (1)$$

그러나 수식 1을 사용하여 토픽 구조를 확률적으로 모두 계산하는 것은 사실상 불가능하기 때문에 깁스샘플링(Gibbs Sampling) 방식을 사용하여 문서 d의 사후확률을 추정하게 된다. 본 연구에서는 [3]를 사용하여 토픽을 추출하였다.

또한 LDA를 사용하여 토픽으로 클러스터링 된 문서들을 파악할 수 있으며, 이를 통해 토픽에 연관된 문서들의 개수를 알 수 있다. 만일 어떤 토픽에 클러스터링 된 문서의 개수가 많다면, 그 토픽은 다른 토픽보다 일반 대중으로부터 많은 관심을 가진다고 볼 수 있다. 따라서 여러 토픽 중에서 중요한 토픽들을 선별적으로 추출할 수 있다.

2.2 토픽 요약

LDA를 사용하여 추출된 토픽은 표 1과 같이 연관어와 토픽 내 연관어의 확률 분포로 구성된다.

단어 (w)	확률 값 (P(w))
은행	0.02
저축	0.018
저축은행	0.012
대출	0.0056
영업	0.0049
정지	0.0045
금융	0.0041
영업정지	0.004
예금	0.003
제일	0.002
...	...

표 1 LDA를 사용하여 얻어진 토픽 예제

표 1의 연관어 관계를 고려하면, 그 토픽은 '저축은행 부실경영'으로 추정할 수 있다. 그러나 이러한 토픽만으로는 그 토픽에 해당하는 구체적이고 세부적인 내용을 알기가 쉽지 않다. 따라서 추출된 토픽과 연관된 세부적인 사건이나 내용을 요약해서 보여주는 것이 필요하다. 본 연구에서는 사용자 후기 게시판 문서들을 문장으로 구분하고, 주어진 토픽과 문장 간의 유사성을 측정하는 목적함수(objective function)를 제안한다. 만일 주어진 토픽에 있는 확률 값이 높은 연관어들이 빈번히 어떤 문장에 포함되어 있으면, 그 문장과 토픽은 유사하다고 가정할 수 있다. 따라서 어떤 문장에 토픽의 단어들이 많이 포함되어 있고 또한 그 단어들의 확률 값의 총합이 높으면, 토픽과 문장은 연관성이 높다고 볼 수 있다. 이러한 가설을 다음과 같은 수식으로 표현된다.

$$P(t|s) = Z(t) \prod_{w \in t} \frac{P(w)}{\sum_{w' \in t \cap sentence(s)} P(w')} \quad (2)$$

수식 2에서 s, t, Z(t)는 각각 문장, 토픽에 속한 단어 집합, 수식 2를 정규화 하는 항을 의미한다. 토픽 t가 입력으로 주어지면, 각 문장 s의 $P(t|s)$을 계산하고, 그 값이 큰 순서대로 상위 k개의 문장을 출력한다.

2.3 토픽 선호 판별

토픽의 긍부정을 판별하기 위해서 먼저 도메인 전문가는 감성사전을 구축한다[1]. 감성사전은 자동차와 관련된 총 540개의 단어로 구성되며, 60%의 긍정 단어와 40%의 부정 단어를 포함한다. 2.1절에서 설명한 LDA를 사용하여 토픽을 추출하고, 각 토픽의 긍부정을 판별하기 위해 다음과 같은 알고리즘을 수행한다.

1) 각 토픽은 확률 값이 큰 순서대로 m개의 연관어를 고려한다.
2) m개의 단어들이 감성사전의 긍부정 단어와 일치하는지를 조사하여 그 토픽에 속한 단어들의 긍정과 부정의 비율을 계산한다.
3) 긍정 비율이 부정 비율보다 높으면 그 토픽은 긍정으로 판단하고, 반대의 경우에는 부정으로 판단한다.
4) 토픽의 긍부정 단어를 사용하여, 2.2절에서 설명한 토픽 요약 알고리즘을 적용하여, 긍정과 부정에 해당하는 문장을 출력한다.

3. 실험 환경 및 결과

3.1 실험 환경

보배드림 사용자 후기 게시판으로부터 총 2,532개의 문서를 수집하였다. 이러한 문서들로부터 토픽을 추출하기 [3]를 사용하였다. 토픽의 개수는 20개로 정했으며, 2.1절에서 설명한바와 같이 토픽에 연관된 문서들의 개수가 많은 순으로 상위 5개의 토픽을 선정하였다. 그리고 토픽 요약 알고리즘을 사용하여 토픽 당 가장 연관성이 높은 문장을 3개씩 선정하였다. 마지막으로 각 토픽에 대한 긍부정을 판단하고, 긍정과 부정에 해당하는 문장을 추출하였다.

3.2 실험 결과

표 2는 보배드림의 사용자들이 가장 관심이 많은 토픽을 나타낸다.

토픽 번호		토픽4	토픽6	토픽10	토픽13	토픽19
순위	1	쏘나타	디자인	연비	가격	자동차
	2	터보	가치	저금리	장착	선정
	3	세단	화제	에코	인하	쏘나타
	4	엔진	안전한	유럽	쏘나타	평가
	5	고성능	신차	가족	휘발유	연비
의미		성능	가치	에코 모드	판매	평가

표 2 소나타 자동차와 관련된 토픽

표 2에서 보는 것처럼, 소나타 자동차에 대해 일반 대중이 관심을 갖는 주요 의제로는 자동차의 성능, 가치, 에코 모드, 판매량, 자동차 평가 등이었다.

표 3은 각 토픽과 가장 관련이 있는 상위 3개의 문장을 나타내며, 각 토픽의 구체적인 정보를 알 수 있다.

토픽	순위	토픽과 가장 연관된 문장
4	1	터보 엔진이 탑재된 쏘나타에는 연비 성능을 대폭 향상시켜주는 변속기인 딘 더블 클러치 트랜스미션도 적용될 예정이다.
	2	터보 엔진 장착 차종이 늘어남에 따라 모터스포츠를 통해 고성능 성격을 홍보하겠다.
	3	터보 엔진을 적용한 쏘나타는 골프의 고성능 모델인 골프의 경쟁모델이 될 것으로 예상된다.
6	1	에드먼즈닷컴 편집장 스캇 올덤은 쏘나타는 차선이탈 경보장치와 전방추돌 경보장치 등 동급 최고의 안전 품목으로 가치가 올랐으며 세련된 디자인과 잘 다듬어진 주행감으로 가족용 세단으로서 손색이 없다고 평가했다.
	2	현대차 관계자는 단순히 가격상승만을 따지기보다는 각종 편의품목에 따른 차의 가치를 고려해야 한다며 구형보다 안전성 상품성 효율이 월등히 높아졌지만 가격인상은 최대한 억제할 것이다.
	3	현대차 관계자는 쏘나타 하이브리드가 인기를 끄는 것은 연비가 뛰어난데다 중고차 가치 보장 프로그램을 도입했다.
10	1	연비 향상에 중점을 둔 에코주행모드를 설정한 후 시동을 껐다 켜도 계속 이 모드가 유지되더군요.
	2	이번 중형 연비 대결은 그간 동호회 내에서 에코 임프레션과 쏘나타 연료효율 비교에 대한 논란에 따라 기획됐다는 게 자동연의 설명이다.
	3	현대자동차는 쏘나타 하이브리드 보유고객을 대상으로 경제운전을 할수록 누적되는 에코 포인트에 따라 혜택을 제공하는 에코 포인트 보상 프로그램을 실시한다고 밝혔다.
13	1	전반적인 내수 침체로 인해 포터와 그랜저를 제외한 나머지 차종은 지난해 대비 판매가 감소했다.
	2	국산차 업계 관계자는 특정 차종을 제외하고 전반적으로 승용차종 판매가 줄어드는 것을 확인할 수 있다며 조만간 차종이 톱에 대거 진입할 것으로 예상한다고 말했다.
	3	현대차는 올 하반기 미국에서 신형 쏘나타 하이브리드와 플러그인하이브리드 모델을 출시하면 판매량이 더욱 증가할 것으로 전망하고 있다.
19	1	현대자동차 신형 쏘나타가 미국 자동차 전문 평가사 오토모티브 사이언스 그룹이하 선정하는 최고의 경제적인 차에 뽑혔다.
	2	컨슈머리포트는 월 자동차 특집호에서 신뢰성 연비 핸들링 승차감 및 공간성 등 종합적인 평가를 통해 쏘나타를 중형부문 최고 차량으로 뽑았다.
	3	현대차 쏘나타가 미국 환경보호청이 발표한 차

		급별 연비 평가에서 위를 차지했다.

표 3 토픽 요약 결과

표 4에서 보는 것처럼, 토픽 긍부정을 수행한 결과, 토픽 6과 10은 중립이었고, 토픽 4와 19는 긍정, 토픽 13은 부정의 의미를 담고 있다.

토픽		토픽과 가장 연관된 문장
4	긍정	자동차업계에 따르면 현대차는 최근 1.6리터 직분사 가솔린 엔진에 터보를 장착한 YF소나타를 준비 중인 것으로 전해졌다.
	부정	
13	긍정	
	부정	디자인에서 호평을 받고 있는 K5 휘발유 모델은 지난달, 올 들어 처음으로 쏘나타 판매량을 앞섰다.
19	긍정	쏘나타의 주행성능과 연료효율을 높게 평가하고 전기차 모드와 하이브리드 모드 간 전환이 부드러워 운전자 피로를 줄였다는 점을 수상 이유로 제시했다.
	부정	

표 4 토픽 선호 조사 결과

끝으로 본 논문에서 제안한 방안의 효용성을 평가하기 위해 군산대학교 통계컴퓨터과학과의 27명의 대학생을 대상으로 사용자 테스트를 수행하였다. 각 평가자에게 토픽 5문항과 토픽 요약 5문항을 제시하고, 각 문항이 유용한지를 물었고, 평가자는 리커트 척도에 의해 '매우 그렇다'는 20점, '그렇다'는 16점, '보통'은 12점, '그렇지 않다'는 8점, '전혀 그렇지 않다'는 4점을 주도록 하였고, 총 점수는 각 100점이 된다. 설문 결과에 의하면 토픽 5문항은 평균 71점, 토픽 요약 5문항은 평균 79점을 기록하여 본 논문에서 제시한 제안 방안이 사용자에게 유익한 정보를 제공한다는 사실을 알 수 있다.

4. 관련연구

조태민과 이지형은 LDA를 이용한 잠재 키워드 추출 방안을 제시하였다. 기존의 키워드 추출 연구들은 문서에서 나타나는 키워드만을 대상으로 하기 때문에, 문서에 직접 등장하지 않는 잠재 키워드를 추출하지 않는 반면, 조태민과 이지형은 잠재 키워드 추출을 위해 주어진 문서와 유사한 문서의 키워드를 후보 키워드로 선택하고 후보 키워드를 구성하는 개별 단어들을 이용하여 후보 키워드의 중요도를 평가하는 방법을 제시하였다[7].

정다미외 5인은 사회적 이슈를 다루는 대용량 뉴스기사를 수집하고 통계적인 기법으로 사회문제를 나타내는 키워드를 추출하는 시스템을 개발하였다. 2009-2012년 동안 국내 10대 주요 언론사에서 생성된 백 30만 건의 뉴스기사로부터 사회문제를 다루는 기사를 식별하고, LDA를 이용하여 사회문제 키워드를 추출하여 웹상에 시계열로 사회문제 키워드의 트렌드를 시각화하여 보여준다[6].

위에서 언급한 방안들은 LDA를 사용하여 잠재 키워드를 추출하는 방안에 초점이 맞추어져 있다. 하지만 본 논문에서는 특정 제품의 평판 마이닝을 위해 토픽 추출 및 요약, 감성분석을 통한 제품의 선호도를 자동으로 출력함으로써 전통적인 여론조사를 대체하는 텍스트 마이닝 방안을 제시한다.

5. 결론 및 향후 연구

본 논문에서는 소나타 자동차의 평판 마이닝을 위한 구체적인 방안을 제시한다. 사용자 후기 게시판으로부터 텍스트를 수집하여 정제하고, 토픽들을 추출한다. 각 토픽은 다수의 고객들이 소나타 자동차에 대해 주로 얘기했던 의제라고 할 수 있다. 이러한 토픽은 서적이나 발표 자료에서 목차에 비유할 수 있으며, 목차의 각 항목의 주요 내용을 찾기 위해 토픽 요약 알고리즘을 사용한다. 예를 들면, 보배드림 사이트에서 주요 의제 중에 하나로 소나타 자동차의 품질이 대두되었다면, 그 토픽과 관련된 구체적인 내용을 요약해서 보여줌으로써 소나타 자동차의 품질은 어떠한지 구체적인 내용을 쉽게 이해할 수 있다. 또한 토픽의 긍부정을 판단하는 알고리즘을 통해 주요 의제들을 긍정과 부정으로 나눌 수 있으며, 각 의제에 대해 긍정적인 점과 부정적인 면에는 구체적으로 어떤 내용이 있는지를 자동으로 파악할 수 있다. 본 연구 결과물은 다양한 제품을 생산하는 산업계에 쉽게 응용 가능하며, 장기적으로 여론조사를 대체할 것으로 예상된다. 신제품의 문제가 무엇인지를 파악하여 제품을 개선하거나, 어떤 방향으로 제품을 홍보할 지에 대한 기초 자료로 활용 가능하다.

향후 연구로는 실험 데이터의 크기를 늘려 소나타 자동차에 대한 심도 있는 분석을 수행할 것이다. 또한 제안 방안을 일반화시킴으로써 다양한 제품에 적용 가능한지를 시험할 것이다. 끝으로 본 연구 결과물을 시연한 프로토타입 시스템을 개발함으로써 기술 수요가 큰 기업체에 적용할 예정이다.

참고문헌

[1] Bing Liu, Sentiment analysis and opinion mining, Morgan&Claypool Publishers, 2012

[2] David Blei, Probabilistic topic models, 55(4):77-84 (2012)

[3] JGibbLDA - A Java implementation of Latent Dirichlet Allocation (LDA), http://jgibblda.sourceforge.net/, 2016

[4] OutWit - Harvest the web, www.outwit.com, 2016

[5] 보배드림, http://www.bobaedream.co.kr/, 2016

[6] 정다미, 김재석, 김기남, 허종욱, 온병원, 강미정, 사회문제 해결형 기술수요 발굴을 위한 키워드 추출 시스템 제안, 지능정보연구 19(3):1-20 (2013)

[7] 조태민, 이지형, LDA 모델을 이용한 잠재 키워드 추출, 한국지능시스템학회 논문지 25(2):180-185 (2015)

[8] 이제 여론조사 시대는 가고... 빅데이터를 요리하라, http://biz.chosun.com/site/data/html_dir/2015/06/05/2015060501615.html?Dep0=twitter, 조선비즈, 2016