한국컴퓨터정보학회 동계학술대회 논문집 제21권 제1호 (2013. 1)

여고생들의 SNS 자료를 이용한 기능성 화장품 기호분석시스템

서정민[*], 송재오[○], 이채리[**], 이상문[***]
[*○](주)디엘정보기술 기술연구소
[**]충주예성여자고등학교
[***]한국교통대학교 컴퓨터정보공학과
e-mail:{jeo, sjm}@dlit.co.kr[○*], bbb0420@naver.com[**], smlee@ut.ac.kr[***]

Functional Cosmetics Trend Analysis System Using SNS Big Data For The Girls High School Students

Jeong Min Seo[*], Jeo Song[○], Chae Ri Lee[**], Sang Moon Lee[***]
[*○]Dept. of Research Center, DLIT Co.
[**]Chungju Yesung Girl's High School
[***]Dept. of Computer Science & Information Engineering, Korea Nat'l Univ. of Transportation

● 요 약 ●

본 논문에서는 사춘기 여고생들의 기능성 화장품의 신상품 개발과 성능 향상을 위한 효율적인 정보의 분석과 생산 정책을 위한 SNS 분석시스템을 제안한다. 제안하는 시스템은 여고생들의 기능성 화장품에 관한 SNS 내용을 분석하기 위한 효율적 알고리즘과 방법론을 제안하여 시스템의 처리량을 최대화하고, 각 작업의 수행시간을 최소화한다. 또한 여고생들의 기능성 화장품에 대한 기호 상태를 파악하여, 그 분석 결과를 제품의 개발 및 생산에 반영하기 위한 비주얼 방법론을 함께 제안한다. 따라서 본 논문에서 제안하는 시스템은 단지 화장품에 대한 분석뿐만 아니라 이와 비슷한 소비자의 기호가 빠르게 변화하는 제조업 분야에서 다양하게 응용이 가능하다.

키워드: 기호변화(Trend), 빅 데이터(Big Data), 소셜 네트워크(SNS)

I. 서 론

현재 우리나라의 제조업은 세계적인 경제위기에서 많은 문제에 직면하고 있다. 이러한 난항을 타개하는 방법으로 제조기술과 S/W 기술을 융합하여 고객의 트렌드를 적시에 정확히 분석하여 제조에 반영함으로써 제품의 가치 상승을 극대화하고 고품질의 제품을 신속하게 개발하여 출시하는 것을 중시하고 있다. 이러한 시스템이 가능하도록 하는 중요한 기술은 소비자의 의견을 조사한 후 전체적인 트렌드의 변화를 분석하는 것이 중요하다. 그러나 불특정 다수의 소비자를 일일이 조사하는 방법은 시간적 · 경제적 비용이 과다하게 요구되므로 현실적으로 불가능하다. 그러나 최근에는 SNS를 이용한 소비자들 간의 활발한 의사소통이 확산되고 있어 이 자료를 이용하면 손쉽게 고객의 기호변화를 분석할 수 있다. 그러나 현재 이러한 자료를 분석하는 시스템 기술이 없어 기업들로는 많은 고민을 하고 있는 것이 현실이다. 이에 본 논문에서는 이러한 문제를 해결하기 위한 SNS 빅 데이터를 이용한 사용자 기호변화 분석시스템을 제안한다.

II. 관련 연구

1. 빅 데이터

빅 데이터의 활용을 위해 주안점을 두어야 하는 부분에 대해 Doug[1]은 빅 데이터의 특징 3가지를 3V로 표현했다. 첫째로 전통적인 데이터 타입과 더불어 새로운 타입의 데이터를 포괄하며 크기(volume)가 방대해 진 것이고, 필연적으로 그에 비례하는 다양성(variety)을 가지고 있으며, 데이터 증가 속도(velocity)가 빠르다는 점이 그것이다. 이러한 요구에 의해 하둡 HDFS 등의 분산 파일 시스템 환경이 주목을 받기 시작했고 단순한 분산 프로그래밍 프레임워크를 넘어 기존 RDBMS의 변형 형태인 NoSQL이 이슈화되기 시작했다. 최근에 급속도로 주목 받고 있는 NoSQL은 Brewer에 의해 제시된 CAP 이론[2] 중 일관성(consistency), 혹은 사용성(availability)보다 분산 수용성(partition tolerance)에 초점을 두어 데이터의 수평적 확장에 용의성에 의의가 있는 데이터 저장/관리 형태이다. Gilbert and Lynch[3]에 의해 증명된 CAP 이론에 따르면 NoSQL은 CAP 3개의 개념을 동시에 충족시킬 수 없다고 했으며, 이에 따라 모든 기존의 RDBMS의 역할을 대신할 수 없는 것은

자명한 사실이다. 비단 NoSQL 뿐만 아니라 어떠한 데이터 분산 저장/관리 형태의 구조라 할지라도 CAP 이론의 제약을 받으므로, 시스템 관리, 구축, 개발에 연루된 모든 사람들은 데이터의 성격과 목적을 분명히 알아야만 빅 데이터에 대해 제대로 인식하고 있다고 할 수 있다. 본 논문의 주요 분석 대상은 SNS자료로 텍스트 마이닝에 밀접한 관련이 있으며 이는 주로 주제추출과 자연언어처리에 근간을 두고 있다. 본 논문에서는 문서자료의 분석을 위해 [4]의 LDA(Latent Dirichlet Allocation) 방법과 [5,6,7] 등을 참조하였다.

2. 하둡의 자료 처리

하둡은 단순한 프로그래밍 모델을 사용하여 대규모의 자료를 분산 처리할 수 있도록 설계된 프레임워크이다[8].

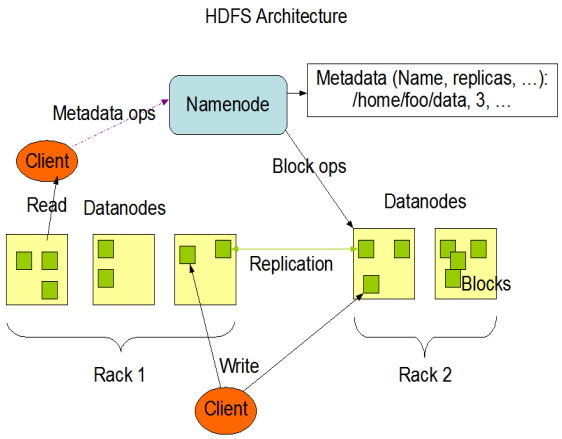

그림 1. 하둡 HDFS 구조
Fig. 1. HDFS Architecture

하둡은 크게 하둡 파일시스템(HDFS)와 맵리듀스(mapreduce)로 구성된다. 하둡 파일시스템은 테라 혹은 페타바이트 규모의 대량의 자료를 저장할 수 있는 분산 파일 시스템으로 고성능의 파일 접근 및 입출력을 제공한다. 이 시스템은 메타 자료를 관리하는 Namenode, 파일시스템의 체크 포인터를 관리하는 Scondarynamenode, 자료가 저장되는 Datanode로 구성된다. 하둡에서 파일은 병렬처리 프로그램을 위한 고가용성을 제공하고 결함에 효율적으로 대처할 수 있도록 미리 정해진 크기(64MB)의 블록 단위로 Datanode에 중복 분산되어 저장된다. 전체적인 동작과정을 그림 1에 나타내었다.

III. 시스템 설계 및 구현

현재 사용하고 있는 SNS는 페이스북을 비롯한 다양한 서비스 모델이 있지만 본 논문에서 분석하고자 하는 대상은 트위터를 기반으로 하였다. 이는 외부에서 접근할 수 있는 트위터 API가 공개되어 있기 때문에 트위터 상에서 다양한 그룹의 의견 및 감정을 실시간으로 수집이 가능하기 때문이다. 따라서 조사하고자 하는 대상 기업의 제품 판매량 및 감정, 사용후기 등을 용이하게 획득이 가능하고 분석하여 의미 있는 분석 결과를 획득할 수 있기 때문이다.

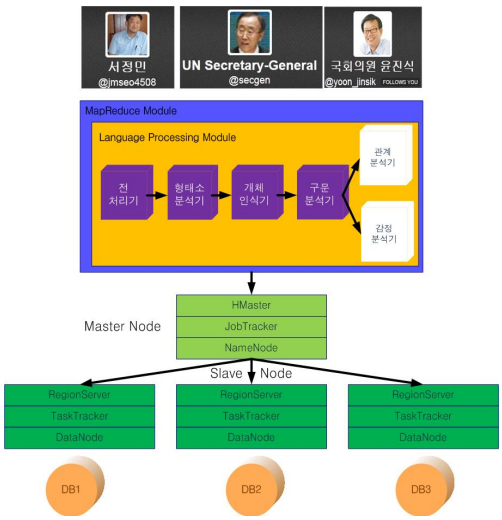

그림 2. 제안 시스템 구조
Fig. 2. Architecture of Proposed System

본 논문에서는 SNS를 분석하기 위해 자연어처리(Natural Language Processing)를 이용한 고객의 오피니언 마이닝 방법을 사용하였다. 따라서 본 논문에서는 본 연구에 참여하는 여고생들의 트위터를 대상 제품의 생산업체의 트위터와 팔로우를 형성하도록 한 후 사용한 화장품에 대한 각종 의견을 수시로 올리도록 하였다. 그림 2는 본 논문에서 구현한 시스템의 전체적인 구조를 보여주고 있다.

시스템은 크게 5개의 모듈로 나눌 수 있는데, 최상위의 모듈은 SNS(Twiter)의 팔로우들로부터 관련 자료를 모으는 Collector 부분이며, 두 번째 모듈은 모은 자료들을 Text-Pre-Processing과 Text-Mining 및 이를 바탕으로 Opinion-Mining을 하는 부분으로 나눌 수 있다. 세 번째 부분은 하둡의 마스터 노드로 본 논문에서는 4개의 서버 중 1개를 HMaster로 지정하였다. 그리고 네 번째와 다섯 번째 부분은 Slave Node와 Database로 구성되어 있는데, 우리는 총 3개의 Slave Node를 구축 하였다. 이는 본 논문에서 구현하고 시험하고자 하는 자료가 일반적인 불특정 다수인을 대상으로 하는 것이 아니고, 특별히 샘플링한 패널들을 대상으로 하기 때문에 그렇게 많은 노드가 필요 없기 때문이다.

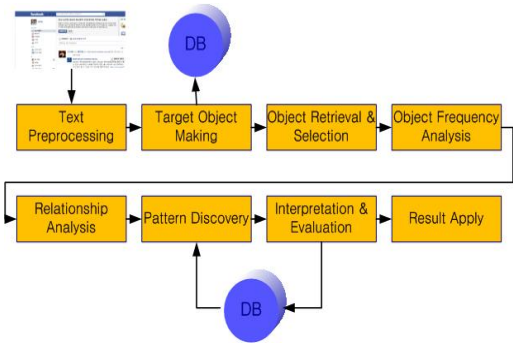

그림 3. 자료의 처리 과정
Fig. 3. Data Processing Logic

그림 3은 제안시스템의 전체적인 자료처리 과정을 보여 주고 있다. 처음 웹봇을 이용하여 페이스북의 자료를 획득한다. 이때 관련 자료를 획득하기 위해 관련 고객들에게 친구를 요청하여 자료의 획득율과 정확율을 높힌다. SNS 자료를 획득한 후에는 불필요한 태그 등을 제외시키는 전처리과정을 거친다. 다음 원하는 제품과 관련된 용어들을 추출한다. 이때 관련된 자료는 미리 DB에 저장한다. 예를 들면 제품의 품명이나 제품과 관련된 단어들을 저장한다. 그리고 추출한 목적 어구들의 빈도와 관련도를 분석한 후 제품에 관한 패턴(기호도 등)을 분석하여 그 결과를 보여주는 과정으로 진행된다. 특히 과거의 분석 결과를 단순히 보여주고 끝내는 것이 아니고 지속적으로 유에 저장하여 과거의 분석 결과와 계속 비교하여 기호도의 변화의 추이를 전체적으로 분석하도록 하였다.

IV. 시험 및 결과 분석

SNS 상의 콘텐츠 및 자료를 분석하면 사회나 조직의 특징 및 행위, 감성 등의 변화나 패턴 등을 발견하여 향후 제품이나 여론의 향방을 예측할 수 있다. 그러나 SNS 기반 예측분석 결과가 정확하려면 예측 대상이 폭 넓어야 하며, 그 내용이 충실해야 한다. 따라서 본 논문에서는 충주의 Y여고생들을 대상으로 충북 진천의 S사의 기능성 화장품을 사용하도록 하고 그 후기를 SNS에 올리도록 하고 그 자료를 분석하였다.

본 논문에서는 상기의 자료를 처리하기 위해 x86 계열 범용 서버 장비 4대로 구성된 하둡 클러스터를 표 1과 같이 구축하였다.

표 1. 시스템 환경
Table 1. System Environment

항목	값
CPU	Intel, 2.5GHz, Quad core
메모리	16GB
HDD	500GB * 2EA
Network	Gigabit Ethernet

그림 3은 실제 자료를 입력하여 실행한 예이다.

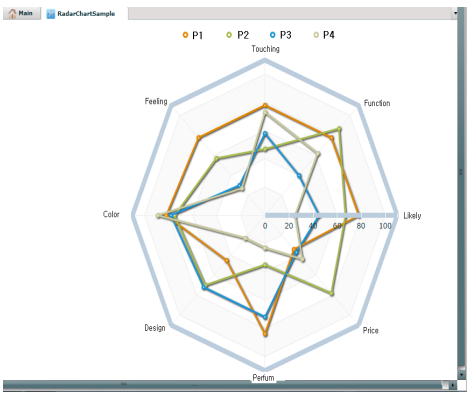

그림 4. 실행 예
Fig. 4. Executed Example

실험에서는 4개의 제품-샴푸, 린스, 바디 클린저, 바디 로션-과 각 제품에 대한 분석을 크게 8종-사용 후 감촉, 기능, 기호도, 가격, 향기, 포장 디자인, 색깔, 심적 느낌-으로 구분하여 분석하였다.

V. 결론

본 논문에서는 하둡을 이용하여 여고생의 기능성 화장품에 관한 SNS 자료를 이용하여 화장품에 관한 기호변화를 분석하였다. 하둡은 대량의 자료 처리에 최적의 환경을 제공하고 있어 본 논문의 연구와 같은 시스템을 구현하는데 적합하다. 따라서 본 논문에서는 하둡을 이용하여 SNS 자료를 수집하여 관련 회사의 제품들에 대한 내용을 텍스트 마이닝 기술을 이용하여 제품에 대한 8가지의 조건을 분석하였다. 그러나 현재의 제안 시스템은 기 결정된 소규모의 패널을 위주로 자료를 수집하고 분석하여 실제 내용과는 정확하게 일치한다고는 할 수 없다. 하지만 표본 집단의 구성이 일정한 연령대로 구성하였으므로 어느정도의 신뢰성을 갖췄다고 믿을 수 있다. 또한 제안시스템을 확장 시 보다 많은 표적집단을 구성할 수 있으므로 향후 그 응용성이 넓다 할 수 있다.

참고문헌

[1] Doug Laney, “3-D Data Management: Controlling Data Volume, Velocity and Variety”, META Group Inc., 2001

[2] Eric A. Brewer, “Towards robust distributed systems”, Proceedings of the nineteenth annual ACM symposium on Principles of distributed computing, pp.4-5, 2000.

[3] Nancy Lynch, Seth Gilbert, “Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services”, ACM SIGACT News, Vol.33, No.2, pp.51-59, 2002.

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, The Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[5] Yung Taek Kim, “Natural Language Processing”, Life & Power Press, 2003.

[6] Sung-sik, Kang, “Han'gugo hyongt'aeso punsok kwa chongbo komsaek”, Hongneung Kwahak Ch'ulp'ansa, 2003.

[7] Young-Min Ahn, Su-Hyun Oh, Yu-Hwan Kang, Young-Hoon Seo, “Answer Extraction of Concept based Question - Answering System”, Vol.2, No.1, pp.448-451, 2005.

[8] Apache Hadoop, http://hadoop.apache.org