JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

응용 및 융합 기술

J. Adv. Navig. Technol. 22(6): 670-674, Dec. 2018

문화권 클러스터링 기반 SNS 빅데이터 및 사용자 선호도 분석

Cultural Region-based Clustering of SNS Big Data and Users Preferences Analysis

노 승 민
성결대학교 미디어소프트웨어학과

Seungmin Rho
Department of Media Software, Sungkyul University, Gyeonggi-do, 430-742, Korea

[요 약]

최근 댓글 / 텍스트, 이미지, 비디오, 블로그 및 사용자 경험을 포함한 소셜네트워크서비스(SNS) 데이터에는 다양한 고객의 추천 시스템을 구축하고 비즈니스 분석가에게 통찰력 있는 데이터 / 결과를 제공하는 데 사용할 수 있는 많은 정보가 포함되어 있다. 멀티미디어 데이터, 특히 이미지 및 비디오와 같은 시각적 데이터는 SNS 데이터 중에서도 특정 (문화권) 지역을 반영할 수 있는 가장 풍부한 데이터이며, 문화적 가치 및 관심사는 전반적으로 데이터의 많은 부분을 차지하고 있다. 이러한 방대한 데이터로부터 원하는 데이터를 지능적으로 추출하고, 엄청난 양의 데이터를 마이닝 하려면 보다 효율적이고 지능적인 데이터 분석 방법이 필요하다. 따라서 본 논문의 목적은 이러한 데이터를 모델링하고, 색인하고, 검색하는 방법에 대해 제안하고자 한다.

[Abstract]

Social network service (SNS) related data including comments/text, images, videos, blogs, and user experiences contain a wealth of information which can be used to build recommendation systems for various clients' and provide insightful data/results to business analysts. Multimedia data, especially visual data like image and videos are the richest source of SNS data which can reflect particular region, and cultures values/interests, form a gigantic portion of the overall data. Mining such huge amounts of data for extracting actionable intelligence require efficient and smart data analysis methods. The purpose of this paper is to focus on this particular modality for devising ways to model, index, and retrieve data as and when desired.

Key word : Cultural clustering, Deep learning, Personal preferences, SNS.

https://doi.org/10.12673/jant.2018.22.6.670

Received 23 October 2018; **Revised** 27 December 2018
Accepted (Publication) 17 December 2018 (30 December 2018)

***Corresponding Author; Seungmin Rho**

Tel: +82-31-467-8348
E-mail: smrho@sungkyul.ac.kr

Ⅰ. 서 론

오늘날 소셜 네트워크 서비스(SNS; social network service)는 전 세계적으로 다국적 문화를 가진 다양한 사람들의 데이터를 나타내는 매우 방대한 데이터 소스이다. 따라서 다양한 유형의 데이터(이미지, 텍스트, 비디오, 가능한 모든 문화 정의 매개 변수 포함)를 통해 여러 형태의 분석을 수행할 수 있는 좋은 데이터 원본이 될 수 있다. 이렇듯 지구상의 모든 문화, 지역에 따른 데이터만 해도 그 규모는 끊임없이 생산되고 있는데 이로 인해 빅데이터를 저장, 처리, 분석하는 프로세스는 정형화가 제대로 되지 않은 상태이다. 따라서 효율적인 분석을 제공하기 위해 지능형 관리 기능을 갖춘 특별한 컴퓨팅 환경과 고급 클러스터 컴퓨팅 기술이 필요하게 된다. 저장 및 분석을 어떻게 하느냐가 대규모 실시간 데이터들을 처리하는 프로세스의 지표가 되기 때문에, 데이터 퓨전 (data fusion)과 딥러닝 (deep learning)의 구현을 위해서 컴퓨팅 파워나 한정된 리소스 관리가 효율적으로 이루어져야 한다 [1].

따라서 본 논문에서는 빅데이터 분석은 데이터 마이닝, 데이터 추출 및 프로세싱을 기반으로 숨은 데이터 간 연관 관계를 찾아내고 다양한 미디어 콘텐츠의 특성 및 다른 문화권에서의 개인 성향 등을 고려한 개인 맞춤형 서비스를 제공하는 것을 목표로 한다. 이러한 빅데이터 기법들은 탁월한 성능을 제공하지만 온/오프라인에서의 위치 정보, 데이터 포맷 변화 및 부적합한 기존의 알고리즘 선정은 빅데이터의 성공적인 분석에 치명적인 영향을 미칠 수 있다.

따라서 빅데이터 수집 및 처리에 특화된 환경 및 시스템이 필요한데 이를테면 wireless ad-hoc과 같이 지속적으로 변하는 센서 네트워크는 위치, 자체적인 데이터 처리, 메모리, 파워 관리, 보안, 로드밸런싱, 확장성과 같은 요소들을 고려해야 하는 것처럼 문화/지역적 군집화 및 대규모 데이터 처리 또한 이에 맞는 적합한 시스템이 필요하다.

현재는 통계적인 관점에서의 빅데이터 분석에 대한 연구가 계속되고 있지만, 이 또한 이러한 대용량 데이터를 표현할 때는 특정한 형식으로 기술되어야 한다. 빅데이터의 통계적 분석은 실시간, 오프라인 도메인을 고려해야하며 데이터는 형태가 불완전하며 다양하기 때문에 어떻게 데이터를 융합(fusion)할 것인가가 중요한 문제로 대두된다. 언급한 통계학 및 기존의 데이터 분석으로는 이러한 문제를 해결할 수 없기 때문에 다양한 관점에서 언급한 문제들을 해결해야만 한다.

본 논문에서는 다양한 미디어 콘텐츠의 특성 및 다른 문화권에서의 개인 성향 등을 고려한 개인 맞춤형 서비스를 제공할 수 있는 추천 프레임워크를 제안하고, 각 세부 모듈별로 구현되어져야 할 기술들에 대해 논하고자 한다.

Ⅱ. 관련 연구

2-1 SNS 메시지 분석을 통한 주요 키워드 추출

최근 몇 년간 트위터, 페이스북, 인스타그램, 링크드인과 같은 다양한 소셜 네트워크 서비스 (SNS; social network service)에서는 메시지 분석을 통한 연관된 키워드 추출을 하는데 관심을 쏟고 있다. 이러한 SNS 데이터들은 기존의 데이터들과 비교해서 짧고 그 형태가 정해져 있지 않다. 이는 SNS 공간에서 사용자들이 상태 공유, 정보 공유, 이벤트 공유 및 다양한 정보를 공유하는데 그 목적이 있고 이를 상업적 형태인 광고로 쓰는 사용자까지 그 용도가 다양하기 때문이다. 따라서, 이러한 메시지 분석을 통해 적절한 키워드를 어떤 사용자에게 어떤 방식으로 제공하며 쓰일 수 있을지에 대한 연구가 최근 급증하고 있다. 특히 링크드인의 경우에는 SNS 서비스를 통해 특정 도메인(구직 및 구인)의 역할을 하기도 한다.

우선 키워드를 추출하기 위해서 여러 알고리즘과 함께 자연어 처리, 분류 등의 연구가 함께 진행되고 있다. 특히 소셜 메시지의 특성상 그 길이가 짧고 형태가 정해져 있지 않기 때문에 이러한 상황을 고려하여 키워드 추출을 하는 것이 중요하며, 데이터의 양과 발생 되는 속도가 페이스북은 1분당 4백만의 포스팅이 될 정도로 기하급수적인 양이기 때문에 대규모 데이터를 효율적으로 처리하는 방법론에 관한 연구도 지속적으로 진행되고 있다.

2-2 멀티미디어 콘텐츠 분석 기술

멀티미디어 콘텐츠 분석은 갈수록 급증하는 멀티미디어 데이터에서 의미있는 콘텐츠를 추출하는 것이 중요하다. 이러한 과정에서 효율적이고 확장적인 방법으로 데이터를 처리해야 하는데 데이터를 처리하는 서버는 어떻게 구현이 되어야 할 것인지, 효율을 증대시키기 위해서 어떤 알고리즘을 써야 할지, 데이터 정제(refinement)는 어떤 식으로 이뤄져야 할지, 의미 있고 필요한 정보 추출을 위해서 데이터 융합 (data fusion) 은 어떻게 해야 하는지 등의 연구가 수행 중이다 [2]. 예를 들어 소셜 미디어 플랫폼에서 빅데이터를 통한 특정 도메인 정보 추천 및 교환 하는 연구가 진행 중이며[3], 대규모 소셜 미디어를 통한 텍스트 및 이미지의 연관 정도를 지표로 삼아 로컬 기반의 서비스를 향상시키는 것에 대한 연구도 진행 중이다.

2-3 딥러닝을 활용한 군집화 기술 및 개인 성향 분석

최근 신경망에 대한 연구가 급부상하면서 딥러닝 (deep learning)에 대한 관심과 그 적용 사례들이 점차 늘어나고 있다. 딥러닝이 가장 큰 강점을 보이는 분야는 패턴인식으로 이미지 및 영상, 음성 인식과 자연어 처리가 대표적이다 [4],[5]. 딥러닝은 방대한 데이터를 분석해 이들의 차이점을 가려내고 유사한 것들을 분류하는 데 강점을 지니고 있다.

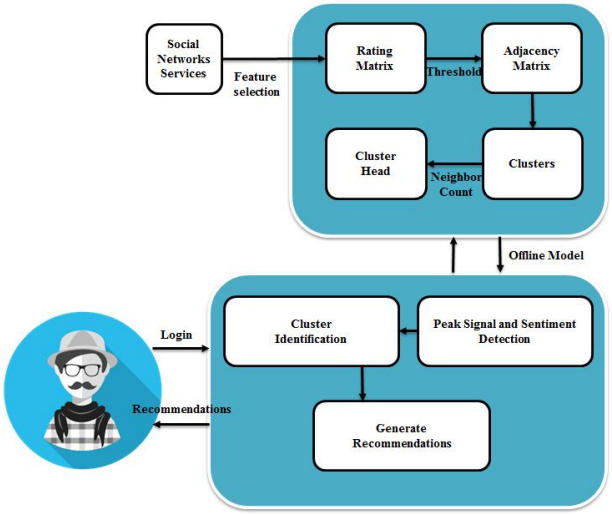

그림 1. 문화권 군집화를 위한 SNS 빅데이터 특성 추출 및 추천 프레임워크

Fig. 1. Framework of the proposed SNS big data feature extraction and analysis for cultural region clustering.

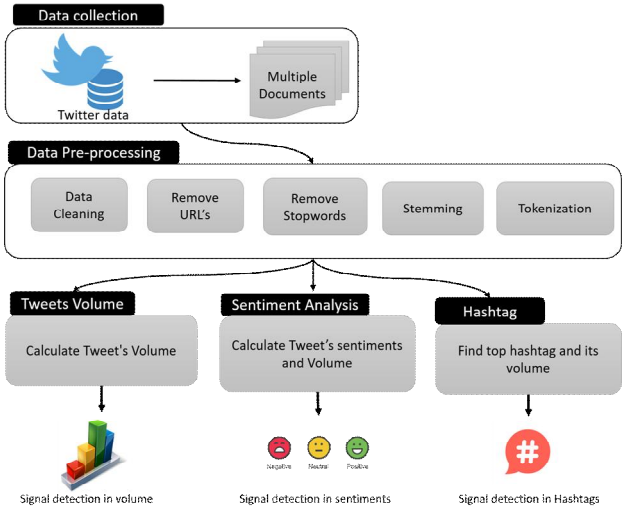

그림 2. 신호/감정 검출 프레임워크

Fig. 2. Framework of the proposed signal detection in social networks.

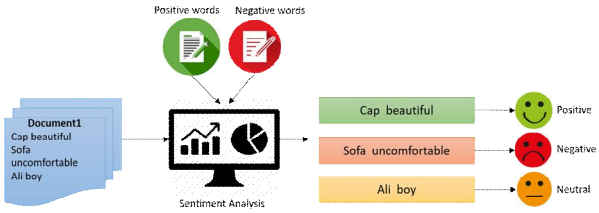

그림 3. 감정분석과정

Fig. 3. Overview of the sentiment analysis.

2-4 멀티미디어 콘텐츠 추천 시스템

대표적으로 IMDB (internet movie database)나 Netflix 서비스는 비디오 콘텐츠를 개인성향에 맞추어 추천하고 있고 나아가 사용자의 위치, 성격, 문화, 환경 등 영향을 미치는 다양한 요소들을 고려하여 추천하는 등의 연구 및 개발이 진행되고 있다. 예를 들면 top-N recommendation을 추출하는 알고리즘을 구현할 때 성능을 어떻게 향상 시킬 것인지 사용자에게 실질적으로 필요한 정보를 제공해주는지에 대한 연구가 끊임없이 진행되고 있다 [6].

III. SNS 빅데이터 기반의 추천 프레임워크

소셜 네트워크 데이터의 양이 매우 방대해지면서 지역적인 트랜드가 다양해지기 때문에 SNS 등의 빅데이터를 정확하게 처리할 수 있는 것은 상당히 중요하며, 딥러닝 기반의 군집화 및 개인 성향 분석이 필요하다. 따라서 본 논문에서는 SNS 빅데이터 특성을 추출하고 분석하여 적절한 멀티미디어 콘텐츠를 제공할 수 있는 프레임워크를 제안한다.

그림 1은 문화권 군집화를 위한 SNS 빅데이터 특성 추출 및 추천 프레임워크로 다음의 2가지 시나리오를 적용할 수 있게 하였다.

- 감정 영향 (sentiment influence): 웹/SNS상에서의 리뷰에서 사용자 관심 분야의 감정 영향을 모델링하였으며, 사용자가 일부 콘텐츠를 사용하기 전에 해당 장소의 콘텐츠 정보를 확인할 뿐 만 아니라 사용자 리뷰에 더 많은 주의를 기울이게 됨
- 알고리즘 및 모델: 사용자 선호도, 지리적 영향, 콘텐츠 효과 및 감정의 영향을 통일된 방식으로 고려하여 다양한 멀티미디어 콘텐츠에 대한 사용자의 좋아하는 행동을 정확하게 포착 할 수 있는 잠재적 클래스 확률 생성 모델이 필요함

그림 2는 SNS에서의 실시간 신호/감정 (signal/sentiment) 검출 프레임워크로 트위터 데이터로부터 전처리과정 (stopword 삭제, 스태밍 등)을 거쳐, 전체 트위터의 볼륨을 구하고 트윗 내의 감정 요소들을 추출하며, 최상위 해쉬 태그들을 검출한다.

감정 분석 과정은 크게 긍정/부정/중립적인 감정으로 분류되며, 트윗(tweet)의 감정을 감지하기 위해 우선 트윗을 단어로 토큰화하고 감정 검출 알고리즘을 토큰에 적용하여 감정을 검출하게 된다. 그림 3은 트윗에서 추출된 감정들의 한 예이다.

최상위 해시 태그를 식별하기 위해 트윗에 있는 모든 해시 태그에 대한 단어 표현이 필요하며, 먼저 모든 해시 태그를 찾은 다음 그림 4와 같이 단어의 집합 (a bag of word)을 얻기 위해 색인을 적용한다 [7].

또한, 사용자의 선호도(preference)를 고려한 향상된 미디어 추천을 해 소셜 멀티미디어의 빅데이터 및 소셜 네트워크 블로

그를 기계 학습 및 딥 컨볼루션 네트워크 (deep convolutional

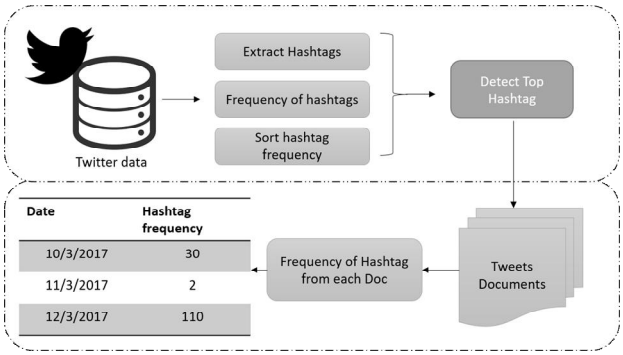

그림 4. 최상위 해시태그 찾는 과정
Fig. 4. Procedure to find top hashtag [7].

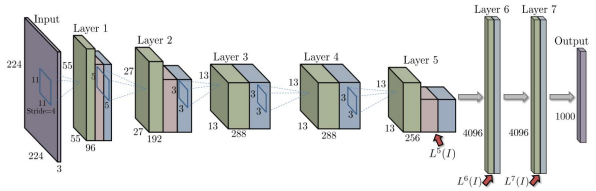

그림 5. 다중 레이어를 가지는 CNN의 구성도
Fig. 5. Convolutional neural network for automatic feature learning.

networks)를 통해 분석하였다. 다음은 소셜 멀티미디어 빅데이터 분석을 위해 사용한 전처리 과정과 중요 특징(feature)들이다.

소셜 네트워크와 온라인 블로그에서 생산/소비되는 거대하고 다양한 데이터는 효과적으로 모델링되기 전에 전처리되어야 하며, 중요한 특성(feature)들을 보다 효율적으로 추출하는데 도움이 된다. 다음은 비주얼 특성(feature) 추출을 위해 사용된 기법들이다.

객체, 사람, 픽셀 등 항목의 saliency(돌출성, 현저함)은 이웃 픽셀/프레임과 비교하여 눈에 띄는 상태를 말하며, 돌출성 탐지(saliency detection)는 시각적 내용 표현을 위한 주요 특징을 식별하는 메카니즘으로 이용하였다 [8]. 이러한 시각적 데이터의 적응 처리는 특징(feature) 추출의 속도 향상 뿐만 아니라 효과적인 표현을 가능하게 해준다.

과적합문제(overfitting issue)를 initialization으로 해결하는 DBN (deep belief network)에 비해 CNN (convolutional neural network)은 모델의 complexity를 줄이는 것으로 해결하기 때문에, CNN과 back propagation 알고리즘을 사용하였으며, 그림 5는 다중 레이어를 가지는 CNN의 구성도를 보여준다.

이 네트워크는 224 x 224 이미지에 적용할 수 있으며, CNN 아키텍처로는 피드 포워드(feed-forward)를 사용하였고, 이미지 I가 주어지면 일련의 레이어 활성화를 생성한다. Gradient vanishing 문제를 해결하기 위해 ReLU (rectified linear unit) 함수를 사용하는데, ReLU 변환 이전에 해당 계층의 활성화(출력)

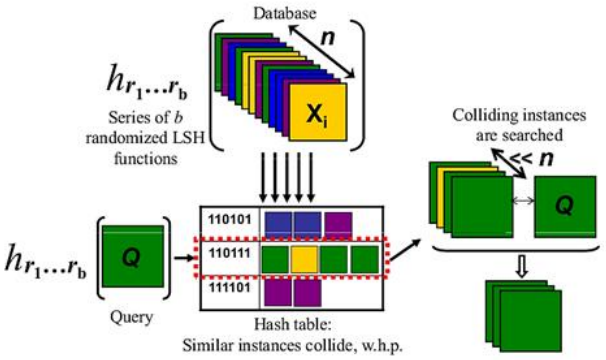

그림 6. 이미지 검색을 위한 LSH 방법
Fig. 6. Illustration of locality sensitive hashing for image searching.

를 L5(I), L6(I) 및 L7(I)로 나타내며 (그림 5), 이들 각각의 고차원 벡터는 입력 이미지의 deep descriptor (nerual code: 신경 코드)를 나타낸다.

분류 알고리즘의 하나인 k-NN (k-nearest neighbor) 알고리즘 로직이 간단하여 구현하기 쉽지만, 학습 모델이 따로 없고, 전체 데이터를 스캔하여 데이터를 분류하기 때문에 데이터의 양이 많아지면 분류 속도가 현저하게 느려진다. 따라서 본 논문에서는 LSH (locality sensitive hashing)를 통해 이 문제를 해결하고자 한다 (그림 6).

Ⅳ. 결 론

본 논문에서는 다양한 문화권에서의 SNS 소비자들의 미디어 소비패턴을 분석하여, 최근 많은 분야에서 활용이 되고 있는 딥러닝 기반의 학습/추천 기법들을 적용한 미디어 빅데이터 및 사용자 선호도 분석을 할 수 있는 프레임워크를 제안하였다.

제안한 시스템은 센서를 통한 행동 패턴 분석 및 SNS를 통한 빅데이터 마이닝 기술, 미디어 추천 시스템을 통해 사용자의 욕구를 더욱 만족시켜 줄 수 있는 서비스로의 활용이 가능하고, 미디어 분석 및 빅 데이터 마이닝, 패턴 분석의 주요 요소 기술들은 지능로봇, 유비쿼터스 응용, 헬스케어 등 최근 부각되고 있는 여러 분야에 광범위하게 적용이 가능하다.

References

[1] G. Abdi, F. Samadzadegan, and P. Reinartz, “Deep learning decision fusion for the classification of urban remote sensing data,” *Journal of Applied Remote Sensing*, Vol. 12, No. 1, 016038, pp. 1-18, Jan. 2018.

[2] J. Popham, M. Forkin, N. Hamblet, and B. Inouye, “Data fusion for sociocultural place understanding using deep learning,” in *Proceeding of the SPIE 10653,*

Next-Generation Analyst VI, 106530E, 27 April 2018.

[3] O. Mastykash, B. Liubinskyi, P. Topylko, and I. Penyak, "Ranking the social media platform user pages using big data," *Mathematical Modeling and Computing*, Vol. 5, No. 1, pp. 56-65, 2018.

[4] J, Ahmad, M. Sajjad, S. Rho, S. Kwon, M.Y. Lee, and S. Baik, "Determining speaker attributes from stress-affected speech in emergency situations with hybrid SVM-DNN architecture," *Multimedia Tools and Applications,* Vol. 77, No. 4, pp. 4883-4907, 2018.

[5] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. Baik, "Action Recognition in Video Sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, Vol. 6, pp. 1155-1166, 2018.

[6] G. Mustafa and I. Frommholz, "Performance comparison of top N recommendation algorithms," in *Proceeding of the Fourth International Conference on Future Generation Communication Technology (FGCT),* Luton, pp. 1-6, 2015.

[7] F. Nazir, M. A. Ghazanfar, M. Maqsood, F. Aadil, S. Rho, and I. Mehmood, "Social media signal detection using tweets volume, hashtag, and sentiment analysis," *Multimedia Tools and Applications,* Online published, pp. 1-34, Aug. 2018.

[8] J. Ahmad, M. Sajjad, I. Mehmood, S. Rho, and S. W. Baik, "Saliency-weighted graphs for efficient visual content description and their applications in real-time image retrieval systems," *Journal of Real-Time Image Processing*, Vol. 13, Issue 3, pp. 431-447, Sep. 2018.

노 승 민 (Rho, Seungmin)
2008년 : 아주대학교 정보통신공학과 (공학박사)
2008년 ~ 2009년 : Carnegie Mellon University, 박사 후 연구원
2009년 ~ 2011년 : 고려대학교 전기전자전파공학부, 연구교수
2012년 ~ 2013년 : 백석대학교 정보통신공학과, 조교수
2013년 ~ 현재 : 성결대학교 미디어소프트웨어학과, 조교수
※ 관심분야 : 음악 검색 및 추천, 집단 및 군집 지성, 시맨틱 웹, 빅데이터