The Journal of the Convergence on Culture Technology (JCCT)
Vol. 2, No. 1, pp.93-98 February 29, 2016. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2016.2.1.93
JCCT 2016-2-9

페이스북 인사이트 데이터 분석

Data Analysis of Facebook Insights

차영준, 이학준*, 정용규**

Young Jun Cha*, Hak Jun Lee, Yong Gyu Jung*****

요 약 최근 정보통신기술의 발달로 인한 각종 모바일 기기와 스마트 기기를 통해 소셜 네트워크 서비스가 많이 대중화 되고 있다. SNS는 오프라인에 존재하는 사회적 관계망이 온라인으로 이동한 친목기반 인맥 형성 서비스이다. SNS는 온라인 커뮤니티와 혼동되어 사용되기도 하지만 차이점이 있다. 이러한 기기들로부터 수집된 정보를 모델링 하는 알고리즘으로는 연관성, 군집화, 신경망, 결정 나무 등의 다양한 기법이 제안되고 있다. 이러한 기법들을 활용하여 여러 가지 방대한 자료를 효과적으로 사용 하는데 연구할 필요가 있다. 따라서 본 논문에서는 특히 군집화에서 좋은 성능으로 평가받는 EM 알고리즘에 대해서 페이스북 인사이트 데이터를 이용하여 군집화를 수행한 결과를 기반으로 알고리즘의 성능을 평가하였다. 이를 통하여 EM알고리즘에 따른 성능의 변화와 남호주 주립도서관 의 실험데이터의 적용결과를 기반으로 분석하였다.

주요어 : 소셜미디어, 관계망, 연관성, EM

Abstract As information technologies are rapidly developed recently, social networking services through a variety of mobile devices and smart screen is becoming popular. SNS is a social networking based services which is online forms from existed offline. SNS can also be used differently which is confused with the online community. A modelling algorithm is a variety of techniques, which are assocoation, clustering, neural networks, and decision trees, etc. By utilizing this technique, it is necessary to study to effectively using the large number of materials. In this paper, we evaluate in particular the performance of the algorithm based on the results of the clustering using Facebook Insights data for the EM algorithm to be evaluated as a good performance in clustering. Through this analysis it was based on the results of the application of the experimental data of the change and the South Australian state library according to the performance of the EM algorithm.

Key Words : Social media, Social networks, Associations, EM

*준회원, 을지대학교 의료IT마케팅학과
**준회원, ㈜innogs, Strategic Planning
**종신회원, 을지대학교 의료IT마케팅학과 (교신저자)
접수일자: 2015년 5월 13일, 수정완료일자: 2015년 7월 6일
게재확정일자: 2015년 8월 20일

Received: 13 May 2015 / Revised: 6 July 2015
Accepted: 20 August 2015
***Corresponding Author: ygjung@eulji.ac.kr
Dept.: Medical IT and Marketing, Eulji Univ.

I. 서 론

최근 정보통신기술의 발달로 인한 각종 모바일기기와 스마트화를 통해 SNS인 소셜 네트워크 서비스가 국내외 많이 대중화 되어 지고 있다. SNS는 오프라인에 존재하는 사회적 관계망이 온라인으로 이동한 친목 기반 인맥 형성 서비스이다. SNS는 온라인 커뮤니티와 혼동되어 사용되기도 하지만 차이점이 있다. 온라인 커뮤니티는 관심사가 비슷한 사람들이 한 장소에 모여 활동하는 그룹 중심의 커뮤니티 서비스인 반면 소셜 네트워크 사이트는 개인이 중심이 되어 관심 있는 다른 개인과 관계를 맺고 더 큰 네트워크를 형성하는 서비스이다. 웹 2.0이등장하면서 나타난 SNS는 오프라인에서만 존재하던 사회적 관계망을 온라인에서 구축한 인맥 형성서비스이다. 지금까지 페이스북, 마이스페이스, 트위터 등과 같은 외국 SNS와 싸이월드, 미투데이 등과 같은 국내 SNS는 지속적 인기를 누리며 인터넷 환경의 주요한 응용분야로 자리잡고 있다. 새로운 커뮤니케이션 채널인 SNS는 사회전반적으로활용되고급속도로확산되고있다. 기존의 카페, 블로그, 싸이월드 등 도 SNS의 한 분류 이지만 태블릿PC, 스마트폰을 활용하여 실시간으로 커뮤니케이션할 수있는 마이크로 블로깅 형태의 사용이 더욱 활성화 되고 있다. 이는 인터넷, 전자상거래 분야는 물론 사회적, 정치적으로도 큰 이슈이며, 인터넷/전자상거래 및 마케팅, 커뮤니케이션 환경의 변화를 가져오고 있다. 이러한 환경과 관심이 반영되어 SNS 서비스에 대한 전망, 앞으로의 활용방안 등에 관한 견해 그리고 시스템 및 사이트 품질, 사용의도, 기타 사회적 문제들과 연계한 다양한 연구가 이루어 지고 있다.

특히 세계적으로 유명한 SNS 서비스인 페이스북은 미국 하버드대학교에서 시작되어 현재는 가장 많은 이용자가 사용하고 있으며 공공기관과 각종 대기업 및 단체에서도 페이스북 페이지를 운영하고 있다. 특히 페이스북은 각종 분야에서 사용자들의 선호도와 시계열 변화에 따른 대중 관심도 분석이 다양하게 이루어 질 정도의 방대한 데이터를 보유하고 있으며, 이를 활용하기 위한 방안이 연구되고 있다. 이러한 다수의 데이터를 기반으로 각 속성 간의 관계 및 특징들을 분석하여 모델링하는 데이터마이닝 분야 또한 최근 들어 각광받고 있다.

이러한 데이터마이닝 분야에서는 데이터에 대한 통계분석이나 모델링을 통하여 정보를 추출해내기 위해서 연관성(Association), 군집화(Clustering), 결정 나무(Decision Tree), 신경망(Neural Network) 등의 다양한 알고리즘들이 연구되고 있는 현황이다. 그러나 실제 문제에서 이러한 기법들을 적용하는 경우, 그 결과에 영향을 미치는 요인이 다수 존재하기 때문에 모든 상황에서 완벽하게 동작할 수 있는 최적의 알고리즘을 선택하는 작업은 단순한 문제가 아니다.

본 연구에서는 군집화 중에서 특히 최단거리를 기반으로 한 K-means 군집화를 사용하지 않고 확률을 기반으로는 군집화 기법인 EM 알고리즘 군집화(EM-algorithm cluster)를 통하여 호주 남부 지역 주립도서관의 페이스북 인사이트 데이터를 분석하고 실험하였다

II. 관련연구

군집화 기법은 패턴인식, 영상처리 및 데이터마이닝 분야에서 매우 유용한 기법으로 사용되어왔다. 최근에는 데이터마이닝 연구에서 다양한 형태의 수치 데이터 및 범주형 데이터로 이루어진 데이터 집합을 처리하기 위해서 많은 연구가 이루어지고 있다. 비록 Gower의 유사도 척도를 이용한 계층적 군집화 기법을 비롯한 다양한 연구가 제안되어져 왔으나, 이러한 기법들은 대규모 범주형 데이터를 포함한 데이터 집합에서는 효율성이 저하된다. 이를 해결하기 위해서 k-means 알고리즘 형태의 기법이 제안되었다. Huang은 표준 k-means 알고리즘을 확장하여 새로운 유사도 척도와 빈도수에 기반을 둔 k-modes 알고리즘을 제안하였다. 또한 k-modes 알고리즘을 Bezdek의 fuzzy c-means 알고리즘의 형태로 일반화한 fuzzy k-modes 알고리즘도 제안하였는데, 실제 데이터 집합에 응용함으로써 그 우월성을 제시하였다. 하지만 대부분의 퍼지 군집화 알고리즘에서, 군집의 중심값은 하나의 스칼라 값으로 표현된다. 이러한 표현 형태는 불확실성과 모호함이 존재하는 공간에서 그 한계점을 지닌다. 따라서 본 논문에서는 EM 알고리즘을 사용하는 군집화 기법을 제안한다.

EM(Expectation maximization) 알고리즘은 1958년 Hartley에 위해서 처음 제안되었고 1977년에 Dem

pster에 의해서 체계화된 군집(Clustering) 알고리즘이다. EM 알고리즘은 K-means 알고리즘과 마찬가지로 초기 모델을 생성한 후 반복 정제과정을 통하여 모델을 최적화된 모델로 만들어간다. EM 알고리즘은 반복 정제 과정을 통하여 각 객체들이 혼합 모델(Mixture Model)에 속할 가능성(Probability)을 조정하여 최적의 모델을 생성해 간다. K-means 알고리즘이 유클리디언(Euclidean) 거리 함수를 사용해서 모델을 생성해 나가는 것과는 다르게 EM 알고리즘은 log-likelihood 함수를 사용하여 모델의 적합성을 평가한다. 즉, EM 알고리즘은 확률기반 군집 (Probability-based clustering)이다.

관측할 수 있는 확률변수 와 관측할 수 없는 확률변수, 그리고 모수 Θ 가 있을 때,(X ,Z)에 대한 확률분포는 $L(\theta;X,Z) = p(X,Z \mid \theta)$로 주어져 있다. 이 때, 최대화 하려는 우도 함수는 다음과 같다.

$$L(\theta;X) = p(X \mid \theta) = \Sigma z p(X, Z \mid \theta)$$

EM 알고리즘은 어떠한 모수 $\theta(t)$를 입력으로 받아서 새로운 모수 $\theta(t+1)$를 찾는 방식인데 이러한 과정이 E(Expectation)과 M(Maximization)단계로 나누어진다. E(Expectation)단계에서는 $\theta(t)$ 가 주어졌을 때 새로운 θ 를 사용할 때 우도의 기댓값 Q를 정의한다.

$$Q(\theta \mid \theta(t)) = E_{Z \mid X, \theta(t)}[\log L(\theta;X,Z)] = \Sigma Z p(Z \mid X, \theta(t)) \log L(\theta;X,Z)$$

M(Maximization)단계에서는 Q를 최대화하는 새로운 모수 $\theta(t+1)$를 계산 한다.

$$\theta(t+1) = \arg\theta \max Q(\theta \mid \theta(t))$$

실제 EM 알고리즘을 사용할 때에는 모수 $\theta(0\})$ 를 적당한 방법(임의값 등)으로 초기화한 다음, $\theta(t)$를 연속적으로 계산하면서 값이 충분히 수렴될 때 멈추는 방식을 사용한다. 관련된 연구로는 기본적인 EM 알고리즘을 조금 변형을 가하여 신호처리 분야에서 사용되는 방법으로써, 기존의 선형적인 클러스터링이 아닌 비선형적으로 경계가 구분되어 Foreground(의미 있는 데이터, signal)와 Background(돌발변수, 의미 없는 데이터, noise)를 구분지어 군집화 하는 Winner-Take-all EM Clustering 방법이 있고, 제한적인 메모리를 가지고 있을 때 대용량의 정보 데이터베이스를 가지고 한 번의 스캔으로 원하는 클러스터링을 할 수 있으며 어느 순간에도 Best Answer를 줄 수 있는 즉, 계산 중에도 solution을 도출할 수 있고, suspendable하고 stoppable하며 resumeable한 EM알고리즘의 개량된 방법이 있다.

III. 실험 및 실험결과

3.1 실험데이터

본 논문에서 사용되어진 남호주 주립 도서관의 페이스북 인사이트 데이터(Facebook Insights Data Export - State Library of South Australia)는 CSV파일로 제공되어 지고 있으며. 남호주 주립 도서관의 페이스북 페이지의 포스팅 된 여러 가지 댓글과 공유 게시글 수, 링크 주소들, 사진 및 비디오등 여러 가지 정보를 가진 데이터 이다. 이 실험데이터는 호주주립 도서관의 속성변수로써 Shares, Userview, Totalview, Likes 등의 7가지 numeric 속성들과 실질적인 리스크를 {Link Photo Video Share} 중의 하나의 값으로 표기하는 nominal 속성인 Type으로 구성되어 있다. 각각의 속성에 대한 세부적인 사항은 아래 표와 같다.

표 1. 실험데이터의 속성

Table 1. Properties of Experimental Data

attribute	type	value
Type	nominal	Link, Photo, Video, Share,
Userview	numeric	continuous from 295 to 3679
Totalview	numeric	continuous from 347 to 4563
Friendstoview	numeric	continuous from 0 to 5786
Likes	numeric	continuous from 1 to 68
Shares	numeric	continuous from 0 to 29
Viewthroghlikes	numeric	continuous from 263 to 1329
LikesnShareuser	numeric	continuous from 0 to 24

3.2 전처리

기본적으로 호주 공공데이터 사이트에서 제공 되어지는 페이스북 인사이트 데이터(Facebook Insights

Data Export – State Library of South Australia) 의 CSV 파일은 불필요한 변수와 속성값들이 많이 존재하고 있다. 전처리 이전 데이터 파일은 다음 그림과 같다.

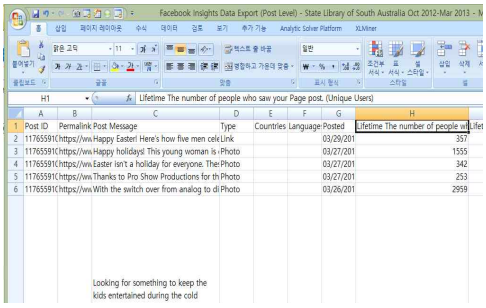

그림 1. 전처리 이전 데이터

Fig. 1 Before Pre-processing Data

원하는 결과값에 불피요한 데이터 속성값들이 존재하는데 전처리 과정을 통해 이러한 속성값들을 삭제하였다. 대표적으로 그림2 스크립트의 좌측의 PostID 속성값과 그 속성에 해당 하는 내용물들을 제거 하였고. 링크 주소를 안내하는 두 번째 변수 또한 제거하였다. 원하는 특정별 군집 형성을 쉽게 구분하기 위해 Type 속성은 제거 하지 않았으며 또 표1 에 나와있는 Attribute 에 해당아지 않는 모든 속성들을 제거하고 변수의 고유 이름도 표1 과 같아 질 수 있도록 변경을 하는 중간단계를 거쳐 전반적으로 실험에 사용되어 질 데이터를 수정을 하는 전처리 과정을 실시하였다. 전처리 후 실험데이터 는 다음 그림 3과 같이 표현되어진다.

Type	Uservieww	totalview	friendstov	likes	shares	viewthrogl	likesnshareuser
Link	357	952	116	4	1	770	4
Photo	1555	3178	2587	68	16	649	13
Photo	342	746	42	6	2	612	4
Photo	253	435	78	26	2	263	24
Photo	2959	6892	5766	52	29	1329	18
Photo	588	1330	313	7	5	928	5
Link	278	625	0	0	0	505	0
Link	352	712	47	6	1	515	6
Link	283	731	0	1	0	602	1
Video	551	1311	117	11	3	1064	9
Link	310	803	2	2	0	642	2
Photo	564	1265	114	12	1	973	11
Link	316	763	5	5	0	606	4
Photo	723	1675	522	21	3	1051	13

그림 2. 전처리 이후 데이터

Fig. 2. After Pre-processing Data

3.3 실험결과

다음의 그림은 Facebook Insights Data Export (Post Level) – State Library of South Australia 데이터를 시각화 하여 나타내므로, 가로축과 세로축의 값은 각 속성값 Attribute에 대하여 안내하며 각각 Type, Uservieww, Totalview, Friendstoview, Likes, Shares, Viewthroghlikes, LikesnShareuser 라는 속성값에 대한 인스턴스 값의 위치를 포인트로 나타내고 있다. 그림을 선택하면 각 속성 변수 값에 대한 결과를 확대하여 볼수 있다. 아래 그림에서 Facebook Insights Data Export (Post Level) – State Library of South Australia 데이터를 시각화 하여 나타낸 그림 중 원하는 X축과 Y축을 지정하여 확대한 것으로, 가로축과 세로축의 값은 각 속성 값 Attribute들 중 원하는 변수를 지정할 수 있고 분산 분포량도 지정할 수 있다.

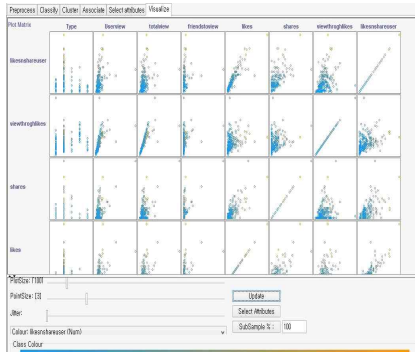

그림 3. 실험결과(1)

Fig. 3. Expeimental Result(1)

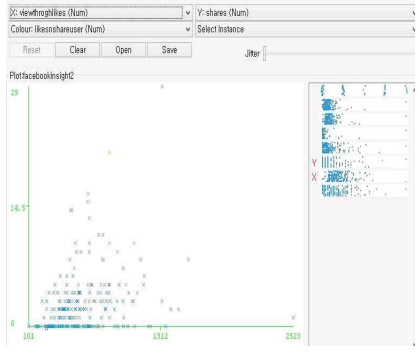

그림 4. 실험결과(2)

Fig. 4. Expeimental Result(2)

IV. 시사점

본 눈문에서는 일반 데이터의 시각화가 아닌 실험데이터를 군집화한 결과를 시각화하기 위하여 아래 그림과 같은 방법으로 출력 값을 표현하였다. 아래 그림은 Weka tool 내에서 EM 군집화를 실행 후 왼쪽 하단 EM 네임명을 오른쪽 마우스로 클릭하면 시각화를 할 수 있도록 하였다. 여기서 Visualize Cluster Assigments를 활용하면 아래 그림과 같은 결과를 도출 할 수 있다.

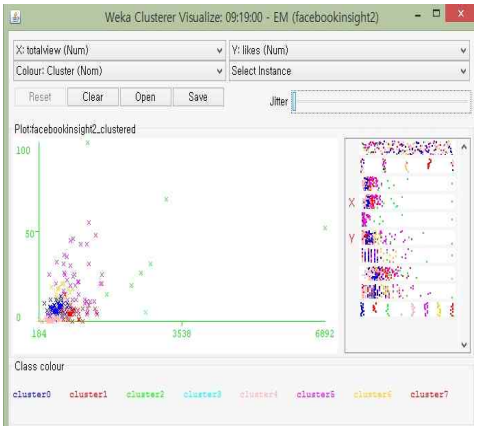

그림 5. EM Cluster 결과화면(1)
Fig. 5. EM Cluster Visualization(1)

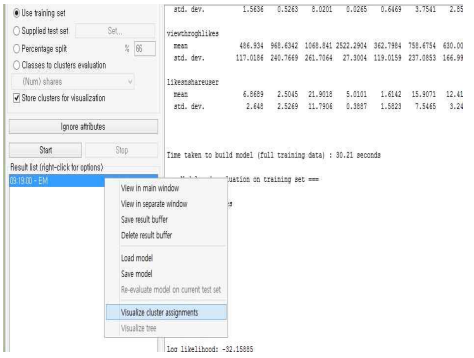

그림 6. EM Cluster 결과화면(2)
Fig. 6. EM Cluster Visualization(2)

예를 들어 조회 수는 없지만, 사람들이 많이 추천한 데이터나, 사람들이 추천하지는 않았지만 수많은 조회수를 자랑하는 특성을 가진 데이터를 군집끼리 분류하여 원하는 데이터에 특성을 파악하여 효과적으로 데이터를 활용할 수 있게 되어 진다. 추후에 실험데이터를 더 많이 수집하고, 이를 기반으로 각 알고리즘을 다양한 관점에서 적용할 필요가 있다.

V. 결 론

최근에는 다양한 분야에서 보유하고 있는 방대한 자료를 기반으로 실질적으로 활용 가능한 정보를 추출하기 위하여 데이터마이닝 분야가 각광받고 있다. 이러한 정보를 모델링하는 알고리즘으로는 연관성, 군집화, 신경망, 결정 나무 등의 다양한 기법이 제안되고 있다. 이러한 기법들을 활용하여 여러 가지 방대한 자료를 효과적으로 사용 하는데 연구할 필요가 있다. 따라서 본 논문에서는 데이터마이닝 기법 중의 하나인 군집화에서 우수한 성능을 자랑하는 EM 알고리즘을 성능을 데이터에 대하여 실험하고 분석하였다.

향후에는 더 활용가능성이 있는 실험을 위하여 실제 데이터를 활용한 사례를 기반으로 성능을 평가하며, 단순한 성능비교뿐만 아니라 bagging, boosting, stacking 등의 앙상블 방법을 통하여 다수의 모델에 복합적으로 적용하는 방향도 연구할 것이다.

References

[1] S. H. Lee. "A study on college classes satisfaction utilizing SNS: Edmodo around the use cases." (2013): 153-169. 이시화. "SNS 를 활용한 대학 수업 만족도에 관한 연구: Edmodo 활용 사례를 중심으로." (2013): 153-169.

[2] H. S. Han and C. I. Kim. "Web accessibility assessment of the social network site." Science of Emotion 12.4 (2009): 481-488. 한혁수, and 김초이. "소셜 네트워크 사이트의 웹 접근성 평가." 감성과학 12.4 (2009): 481-488.

[3] E. S. Lee and Y. S. Lim. "The message structure analysis with exploratory study refers to marketing communications networks in the domestic company to take advantage of Facebook." Korea Advertising Gazette 14.3 (2012): 124-155. 이은선, and 임연수. "페이스북을 활용한 국내 기업의 마케팅 커뮤니케이션에 대한 탐색적 연구 의미연결망을 통한 메시지 구조 분

석." 한국광고홍보학보 14.3 (2012): 124-155.

[4] J. H. Du and J. H. Kim. "Effects of the Facebook ad types." Korea Advertising Gazette 14.2 (2012): 300-330. 두진희, and 김정현. "페이스북 광고 유형에 따른 효과 연구." 한국광고홍보학보 14.2 (2012): 300-330.

[5] S. S. Lee "A Preliminary Study on the library's Facebook page actual conditions." South Korea 43.4 Library and Information Science (2012): 347-372. 이수상. "도서관 페이스북 페이지의 운영 실태에 관한 기초연구." 한국도서관·정보학회지 43.4 (2012): 347-372.

[6] J. A. Seol. "Study on the use of Facebook and privacy." Media and law 11.1 (2012): 63-92. 설진아. "페이스북 이용과 프라이버시 침해에 관한 연구." 언론과법 11.1 (2012): 63-92.

[7] D. W. Kim and K. H. Lee. "A Fuzzy Clustering Algorithm for Clustering Categorical Data." Journal of Korean Institute of Intelligent Systems 13.6 (2003): 661-666.

[8] J. W. Kim "Improved Artificial Intelligence class with WEKA tool." Proceedings of KIIS Fall Conference. Vol. 22. No. 2.2012. 김종완. "WEKA 도구를 이용한 인공지능 수업 개선." Proceedings of KIIS Fall Conference. Vol. 22. No. 2. 2012.