Journal of the Korea Convergence Society
Vol. 8. No. 8, pp. 71-75, 2017
https://doi.org/10.15207/JKCS.2017.8.8.071

키워드를 위한 시퀀셜 패턴 평가 지표와 SNS 팔로워의 관계를 이용한 사용자 관심사항 추출방법

신봉희[1], 전혜경[2*]
[1]인천대학교 컴퓨터공학부, [2]인하대학교 컴퓨터공학부

Extracting Method of User's Interests by Using SNS Follower's Relationship and Sequential Pattern Evaluation Indices for Keyword

Bong-Hi Shin[1], Hye-Kyoung Jeon[2*]
[1]Dept. of Computer Science & Engineering, Incheon National University
[2]Dept. of Computer Science & Information Technology, Inha University

요 약 SNS 등의 보급으로 인해 Web 기반의 소비자 생성 데이터는 기하급수적으로 늘어나는 추세이다. 수많은 데이터 속에서 사용자의 관심에 맞는 콘텐츠를 정확히 추출하는 것은 여러 분야에서 중요하다. 특히 비즈니스 분야에서는 많은 사용자들 속에서 자신들에게 적합한 고객을 찾아 마케팅 정책을 수립하는 것이 중요하다. 본 논문에서는 트위터의 팔로우-팔로잉 관계를 통해 각 계정에 관심이 있는 고객들을 중심으로 중요한 정보를 얻고자 한다. 현재 트위터의 팔로워 간의 관계는 사용자의 세부 관심 사항을 반영하지 않는다. 그러므로 본 연구에서는 팔로우들의 트윗에 대한 키워드 추출 방법을 사용하여 세부 관심 사항을 파악하려고 한다. 이를 위해 국내 상업 트위터 계정 2곳을 선정하여 팔로워로부터 수집한 텍스트 데이터의 마이닝 핵심 문구에 대한 순차 패턴 평가 지표를 적용한다.

• 주제어 : 융합, SNS, 정보검색, 단어빈도, 문서빈도, 확신도

Abstract Due to the spread of SNS, web-based consumer-generated data is increasing exponentially.It is important in many fields to accurately extract what is appropriate for the user's interest in a large amount of data. It is especially important for business mangers to establish marketing policies to find the right customers for them in many users. In this paper, we try to obtain important information centering on customers who are interested in each account through Twitter follow - following relationship. Because Twitter's current follower relationships do not reflect the user's interests, we try to figure out the details of interest using keyword extraction methods for tweets of followers. To do this, we select two domestic commercial Twitter accounts and apply the sequential pattern evaluation index to the mining key phrase of the text data collected from the follower.

• Key Words : Convergence, SNS, Information Searching, TF, DF, Confidence

*Corresponding Author : 전혜경(jhk7010@nate.com)
Received July 10, 2017 Revised August 9, 2017
Accepted August 20, 2017 Published August 28, 2017

1. 서론

SNS(Social Network Service) 등의 보급에 의해 Web 기반의 소비자 생성형 콘텐츠로서 수없이 생성되는 다양한 형식의 데이터의 용량은 갈수록 증가하고 있다. 많은 BtoC 기업은 빅데이터(Big Data)로 부르는 이러한 대량의 데이터에서 소비자의 동향에 대한 통찰력을 얻을 수 있는 기회를 얻고자 한다. 한편, 소비자의 잠재적인 흥미와 관심이 무엇인지 정확하게 파악하는 것은 어렵고 전문적인 지식과 기술의 획득이 필요하기 때문에 효과적인 지원 방법이 필요하다[1,2]. 특히 텍스트 데이터에서 통계적인 성질을 이용하여 키워드와 텍스트의 유사성에서 사용자의 흥미와 관심을 이해하는 유용한 정보를 얻고자하는 텍스트 마이닝 기술은 효과적인 지원 방법의 하나로서 기대되고 한다. 본 연구에서는 Twitter에서 소비자의 관심을 나타내는 언어적으로 공개된 특징적인 단어와 절을 꺼내기 위해서, Twitter에서 팔로우하는 행위에 주목했다. 이들은 계정을 팔로우하는 사용자는 비슷한 흥미와 관심을 가지고 능동적으로 그 흥미와 관심을 팔로우 행위로 나타낸 것으로 가정하고 있다. 그러나 팔로우하는 행위만으로는 흥미와 관심으로 이어질 특징적인 단어를 파악할 수 없기 때문에 사용자간에 동일하게 나타나는 특징 단어의 식별을 할 필요가 있다. 본 논문에서는 기존 단어의 출현 빈도에 따른 중요도 지표 이외에 어구를 단어의 계열 패턴으로 한 계열 패턴 지표의 정의 및 산출 결과의 비교에 국내에서 기업 공식 계정을 팔로우하고 있는 텍스트 데이터 비교를 실시한다. 이 결과를 바탕으로 각 공식 계정의 팔로워가 관심을 가지는 특징적인 단어의 추출과 각 지표의 관련성에 대해 고찰한다[3,4,5].

2. 관련연구

트위터 등 SNS로써의 성질을 갖는 서비스에서는 사용자 간의 관심을 계정 간의 네트워크 구조로 얻을 수 있다. 네트워크의 국소적인 구조는 사용자 계정을 node로, 사용자간의 관계 (팔로잉이나 친구) 등을 edge로 표현할 수 있다. 이러한 네트워크 구조에 대해 HITS(Hyperlink-Induced Topic Search) 알고리즘에 의한 PageRank와 중심성의 각 지표를 사용하여 네트워크의 중심을 계량화하여 정의한다. 사용자의 관심은 중심에 가까운 허브라는 노드 친밀감과 관련된 에지 개수 (차수)로 나타낼 수 있다 [6]. Fig.1은 HITS 알고리즘의 의사코드이다[7].

```
1 G := set of pages
2 for each page p in G do
3   p.auth = 1 // p.auth is the authority score of the page p
4   p.hub = 1 // p.hub is the hub score of the page p
5 function HubsAndAuthorities(G)
6   for step from 1 to k do // run the algorithm for k steps
7     norm = 0
8     for each page p in G do  // update all authority values first
9       p.auth = 0
10      for each page q in p.incomingNeighbors do // p.incomingNeighbors is the set of pages that link to p
11        p.auth += q.hub
12      norm += square(p.auth) // calculate the sum of the squared auth values to normalise
13    norm = sqrt(norm)
14    for each page p in G do  // update the auth scores
15      p.auth = p.auth / norm  // normalise the auth values
16    norm = 0
17    for each page p in G do  // then update all hub values
18      p.hub = 0
19      for each page r in p.outgoingNeighbors do // p.outgoingNeighbors is the set of pages that p links to
20        p.hub += r.auth
21      norm += square(p.hub) // calculate the sum of the squared hub values to normalise
22    norm = sqrt(norm)
23    for each page p in G do  // then update all hub values
24      p.hub = p.hub / norm   // normalise the hub values
```

[Fig. 1] HITS Algorithm

한편, SNS에서는 텍스트 데이터로 사용자의 정보가 발생되며 그 내용은 주로 사용자의 흥미와 관심을 나타낸다. 텍스트 데이터에서 사용자의 흥미와 관심을 나타내는 같은 특징적인 어구의 추출에는 어구의 출현 빈도에 따른 다양한 지표의 이용을 생각할 수 있다. 이러한 지표 중 공통적으로 발생되는 정도를 기준으로 단어 사이의 네트워크 구조를 얻음으로써 시각적으로 흥미와 관심도가 높은 단어 및 어구의 추출을 지원하는 방법에 대해서도 일반 텍스트 마이닝 도구 [8] 등을 통해 이용 가능하다. 그러나 의미 없는 신변잡기적인 데이터에 대해 어떤 전처리가 필요하다던가, 단어의 중요도 등을 경량화하기 위해서 어떠한 지표가 타당한지를 결정하는 것은 적용분야에 따라 크게 다르기 때문에 아직 많은 검토가 필요하다. 따라서 사용자의 흥미와 관심 등이 분석가의 시각적 감각에 맞는 중요 단어를 계량적으로 나타내는 지표 개발이 필요하다고 생각된다[9].

3. 텍스트 내 어구의 중요도평가 지표

단어 또는 여러 단어로 구성된 어구의 중요도를 측정하는 자연어 처리 및 텍스트 마이닝은 다양한 중요도 평가 지표가 개발되어 왔다. 이러한 지표의 기준이 되는 것

은 단어의 출현 빈도이며, 단어의 출현 빈도는 단어가 출현한 횟수를 나타내는 단어 빈도 (TF : term frequency)와 어구가 포함된 문서 수에 해당하는 문서 빈도 (DF : document frequency)의 2개의 출현 빈도 측정 기준이 일반적으로 사용된다[10,11,12].

Table 1는 l개의 단어로 구성된 어구 $term_i = < w_{i,...,} w_l > (l \geq 1)$의 대표적인 평가 지표를 나타낸다. TFIDF는 TF와 DF 모두를 고려하고 TF 대상 전체 문서 $|D|$와 DF의 비율을 가중치로 한 중요도 지표로서 중요 단어 추출에 잘 사용된다. 또한 2 개의 출현 빈도 측정 기준에 대해 간단한 비율로 출현 빈도에 따른 특성을 측정하기 위한 지표를 정의 할 수 있다.

<Table 1> The significance indices according to the occurrence frequency measure for words

	Frequency measurement Indices of words $term_i$	
	$DF = \|D_{\in term_i}\|$	$TF = \sum_j freq(term_i, d_j)$
Support	$DF/\|D\|$	$TF/\sum TF_{term_i}$
Odds	$DF/(DF-\|D\|)$	$TF/(TF-\sum TF_{term_i})$
Entropy	$(DF/\|D\|)\log_2(DF/\|D\|)$	$(TF/\sum TF_{term_i})\log_2(TF/\sum TF_{term_i})$
Jaccard coefficient	$\frac{DF}{DF(w_1 \cup .. \cup w_l)}$	$\frac{TF}{TF(w_1 \cup ... \cup w_l)}$
TFIDF	$TF*\log(\|D\|/DF)$	

4. 계열 패턴 지표

계열 패턴의 평가 지표는 아이템 집합 I에 속하는 아이템 $i \in I$ 인 계열 데이터 $s_i = <i_1, ..., i_m>$로 이루어진 계열 데이터 세트 $D = s_i$의 부분 계열 α의 출현 빈도 $freq(\alpha, D)$에 따라 계열 패턴의 다양한 특성을 계량화하는 지표이다. 계열 데이터 세트 D의 부분 계열 $\alpha = <i_1, ..., i_j> (j < m)$는 출현 빈도의 나타내는 중요한 평가 지표로 사용되는 대표적 기준 중 하나이다. 각 문서의 중복을 고려한 빈도 기준 TF와 중복을 고려하지 않은 빈도 기준 DF를 사용하고, 순서가 없는 아이템 집합 평가 지표군과 계열 패턴의 항목을 고려한 평가 지표군을 기초로 하여, 확신도를 기반으로 한 지표를 정의한다[13]. 여기에서 α를 어구 $term_i$, 각 아이템을 어구 $term_i$ 안의 단어 w_k로 한다.

<Table 2> The sequence pattern indices by frequency of occurrence measure for partial sequential α

	Frequency measurement Indices of words $term_i$	
	$DF = \|D_{\in term_i}\|$	$TF = \sum_j freq(term_i, d_j)$
Heading Confidence (H-Conf)	$DF/DF(w_1, D)$	$TF/TF(w_1, D)$
Max Confidence (MaxConf)	$\max\left(\frac{DF}{DF(w_k, D)}\right)$	$\max\left(\frac{TF}{TF(w_k, D)}\right)$
All Confidence (AllConf)	$\frac{DF}{\max(DF(w_k, D))}$	$\frac{TF}{\max(TF(w_k, D))}$
Sequential Confidence (SeqAllConf)	$\frac{DF}{\max(DF(\beta_{\subseteq term_i}, D))}$	$\frac{TF}{\max(TF(\beta_{\subseteq term_i}, D))}$

Table 2에 나타낸 바와 같이, 계열 패턴으로 항목 간 순서 관계를 고려하면 α의 출현 빈도와 α의 부분 계열 β 출현 빈도와 비율인 신뢰도에 대해 8가지 지표를 정의할 수 있다.

5. 평가지표별 팔로잉 관련어구 비교

이번 절에서는 Twitter에서 제공하는 WebAPI[14]를 이용하여 얻은 유명 트위터 계정의 데이터에 대한 어구의 중요도 평가 지표군, 계열 패턴 지표 군에 의한 특징적인 단어를 추출한다. 각 계정의 팔로워의 차이에 따라, 추출되는 용어의 차이 및 각 평가 지표의 정렬 결과를 비교한다.

5.1 텍스트 데이터 수집 및 자동 용어 추출 방법에 의한 후보 용어의 구분

본 실험에서는 인터파크티켓, 예스24의 각 공식 계정에 대해 각 계정의 팔로워 텍스트 집합을 각각 평가 지표 산출을 위한 텍스트 집합으로 한다.

각 텍스트 집합의 수집은 2017 년 5월 22일부터 5월 26일 5일 동안 각 계정의 팔로워 당 최대 5000 개의 사용자 id 대해 각각 최근 100 트윗을 상한으로 설정하여 실시하였다.

다음으로 수집된 텍스트 집합에서 특징적인 단어가 될 수 있는 후보 단어를 얻기 위해 자연 언어에서 사용되는 자동 용어 추출 방법을 각 텍스트 집합에 적용한다. 본 실험에서는 FLR 점수를 기반으로 한 자동 용어 추출

방법 [15]을 이용했다. 이 때, FLR 점수 산출에 필요한 명사의 분류는 MeCab [16]과 동시에 배포되는 IPA 사전(mecab-ipadic-2.7.0-20070801)을 이용했다. $FLR(term_i, D) > 1$ 인 후보 계열 패턴과 각 계정의 팔로워 의한 텍스트 집합과 각 텍스트 집합에서 자동 용어 추출 결과를 Table 3에 나타낸다.

<Table 3> Numbers of Collected Data

Account	Numbers of Followings	$\|D\|$	Candidate set
Interpark	4257	337499	296171
Yes24	4321	351841	283457

5.2 평가지표군별 결과 비교

이 절에서는 각 평가 지표군의 평가지표를 이용하여 선택한 특징적인 단어를 각 평가 지표의 정렬 결과와 비교한다. 두 공식 계정의 팔로워 트윗에서 얻은 후보 단어 중 *FLR*점수가 상위 1000 개인 단어에 대한 문헌빈도 *DF*, *TFIDF*, 문헌빈도의 확신도, 문헌 빈도에 의한 Jaccard 계수, 문헌빈도에 의한 선두확신도의 각 값을 문헌빈도에 의한 Jaccard 계수로 정렬하여 표 4 표 5에 나타낸다. 그러나 하나의 단어로 구성되어 단어 빈도가 100을 밑도는 단어는 대상에서 제외했다. 그 결과 많은 이모티콘 표현이 상위를 차지하게 되는 특징은 보였다.

<Table 4> Results by index on Yes24's Twitter

Word	TF	TFIDF	H-Conf(TF)
문재인대통령	684	9889.26	0.96
19대 대선	415	8127.43	0.93
문재인	912	3271.74	0.93
탄핵	735	2716.58	0.90
배송	501	1921.35	0.89
ㅂㄷㅂㄷ	1751	4299.03	0.74
ㅋㅋㅋㅋ	3544	1928.30	0.53
이벤트	2333	5165.97	0.43
구매	751	5693.33	0.41
할인	860	1872.73	0.32

<Table 5> Results by index on Interpark's Twitter

Word	TF	TFIDF	H-Conf(TF)
잠실실내체육관	115	1974.76	0.81
트와이스	2435	914.85	0.67
티켓팅	2176	3214.25	0.49
옥주현	1248	2678.57	0.37
정택운	1972	2502.27	0.37
ㅋㅋㅋ	1283	1478.58	0.28
!!!	2161	2876.38	0.26
^^	3811	3774.32	0.19
켄	257	1433.55	0.17
서울재즈페스티벌	2541	2064.89	0.15

인터파크 티켓에서는 급등 연관어에 아이돌의 이름이 많이 추출되었다. 인터파크의 트위터는 카테고리별로 트위터를 운영 중인데 공연정보가 주로 제공되는 특성상 아이돌이 출연하는 공연에 많은 관심을 사용자들이 보인 것으로 생각된다. 또한 예스 24는 연관어에 문재인과 타임지가 등장했는데 이는 5월 9일 실시한 대선 영향에 따른 시기적 특성에 따른 것으로 생각된다.

6. 결론

본 논문에서는 Twitter에서 많은 팔로워를 가진 계정에 주목했다. 팔로워가 생성하는 텍스트 데이터 에서 팔로워의 흥미와 관심에 관련된 특징적인 단어의 추출을 위해 단어의 중요도 지표 및 계열 패턴 평가 지표의 비교했다. 그 결과 어구 전체 출현 빈도뿐만 아니라, 어구의 구성 요소인 단어 (일부는 기호인 이모티콘)를 고려 지표를 이용함으로써 더 특징적인 단어를 선택하는 것을 보여줄 수 있었다. 그러나 이모티콘 삽입 등 텍스트에 의한 감정 표현이 어떤 관심의 반영을 반영 할 것인지 대해서는 좀 더 연구해 볼 가치가 있을 듯 하다.

향후 연구로는 위키피디아의 기사 제목을 사용해서 고유 명사 사전을 구축하여 보다 정확한 후보 단어의 추출하고자 한다. 또한 계열 패턴 평가 지표에 대해서는 자주 등장하는 아이템 집합의 평가 값을 이용하여 다양한 지표로 계열 패턴의 아이템 간의 순서 관계에 맞게 책정할 수 있도록 한다. 그리고 특징적인 단어의 추출이 가능한 지표의 개발을 목표로 한다.

ACKNOWLEDGMENTS

본 논문은 인천대학교 2016년도 자체연구비 지원에 의하여 연구되었음.

REFERENCES

[1] M. S. Lee, “A Study on Characteristics of Eco-friendly Behaviors using Big Data: Focusing on the Customer Sales Data of Green CardStudy of GUI design convergence”, Journal of Digital Convergence, Vol. 14, No. 1, pp. 151-161, 2016.

[2] D. I. Tak, “A Study on The Inluence of Convergence Benefit of Facebook Fan Page in Brand Attachment and Brand Commitment,” Journal of the Korea Convergence Society , Vol. 6, No. 9, pp. 199-206, 2015.

[3] S. W. Lee, S. H. Kim, “Finding Industries for Big Data Usage on the Basis of AHP”, Journal of digital Convergence, Vol. 14, No. 7, pp. 21-27, 2016.

[4] M. H. Lee, “Vitalizations of Government Information Sharing, System Connection, and System Integration”, The Korea Institute of Public Administration, 2014.

[5] S. K. Park, “Proposal of a mobility management scheme for sensor nodes in IoT(Internet of Things)”, Journal of Convergence Society for SMB, Vol. 6, No. 4, pp. 59-64, Dec. 2017

[6] J. M. Kleinberg, “Authoriatative Sources in a Hyperlinked Environment”, Journal of ACM, Vol, 46, Issue, 5, pp. 604-632, 1999

[7] http://em.wikipedia.org/wiki/HITS_algorithm

[8] J. Franke, G. Nakhaeizadeh, I. Renz, "Text Mining", Theoretical Aspects and Applications. Heidelberg: Physica-Verlag. 1-19. 2003

[9] L. S. Kim, “Convergence of Information Technology and Corporate Strategy”, Journal of the Korea Convergence Society, Vol. 6, No. 6, pp. 17-26, 2015.

[10] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques 3rd Edition, 2011.

[11] Y. J. Kim, “Convergence of Business Information System Process using Knowledge- based Method”, Journal of the Korea Convergence Society, Vol. 6, No. 4, pp. 65-71, 2015.

[12] M. H. Lee, “A Study on N-Screen Convergence Application with Mobile WebApp Environment”, Journal of the Korea Convergence Society, Vol. 6, No. 2, pp. 43-48, 2015.

[13] T. Wu, Y. Chen, J. Han,"Association Mining in Large Databases: A Re-examination of Its Measures", In Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 621-628 (2007)

[14] https://dev.twitter.com/resources/twitter-libraries

[15] H. Nakagawa, "Automatic term recognition based on statistics of compound nouns", Terminology, Vol.6,No.2, pp.195-210 (2000)

[16] https://code.google.com/p/mecab/

저자소개

신 봉 희(Bong-Hi Shin) [정회원]

- 1977년 인하대학교 전자공학과 공학사
- 1981년 인하대학교 전자공학과 공학석사
- 1995년 단국대학교 전자공학과 공학박사
- 2010년~현재 인천대학교 컴퓨터공학부 교수

<관심분야> : 마이크프로 프로세서, 임베디드시스템, 사물인터

전 혜 경(Hye-Kyoung Jeon) [정회원]

- 1995년 2월 : 인하대학교 일문과 (문학사)
- 1999년 8월 : 인하대학교 정보공학과(공학석사)
- 2002년 9월 ~ 2005년 8월 : 인하대학교 컴퓨터정보학과 박사수료
- 2009년 4월 ~ 2015년 2월 : 이스트림 선임연구원
- 2016년 3월 ~ 현재 : YM나을텍 선임연구원

<관심분야> : 상황인식, 센서네트워크, 유비쿼터스, 사물인터넷