ISSN 1598-0170 (Print)
ISSN 2287-1136 (Onine)
http://www.jksii.or.kr

트위터를 이용한 이벤트 감지 시스템☆

Event Detection System Using Twitter Data

박 태 수[1] 정 옥 란*
Tae Soo Park Ok-Ran Jeong

요 약

최근 소셜 네트워크 사용자들이 늘어나면서, 각 지역에서 관심 받고 있는 사회적인 이슈나 재해 등과 같은 이벤트에 대한 정보들이 소셜 미디어 사이트를 통해 실시간으로 빠르게 대량으로 게시되고 있으며, 사회적 파급효과도 매우 커지고 있다. 본 논문에서는 지역정보를 가진 트위터 데이터를 이용하여 특정 시간, 지역에 사용자들이 관심을 가지고 있는 이벤트를 탐지하는 방법을 제안하고자 한다. 이를 위해 트위터 스트리밍 API를 이용해 데이터를 수집하고, 트윗의 키워드들의 시간에 따른 빈도수를 분석하여 정상적인 패턴과 다른 패턴을 가진 키워드를 이벤트로 추출하고, 같은 이벤트에 대한 키워드들을 군집화 하기 위해 co-occurrence 그래프를 이용하여 이벤트 감지 시스템을 구현하였다. 그리고 실험을 통해 제안한 기법의 유효성을 검증한다.

☞ 주제어 : 이벤트 감지, 소셜 네트워크, 소셜 미디어 콘텐츠

ABSTRACT

As the number of social network users increases, the information on event such as social issues and disasters receiving attention in each region is promptly posted by the bucket through social media site in real time, and its social ripple effect becomes huge. This study proposes a detection method of events that draw attention from users in specific region at specific time by using twitter data with regional information. In order to collect Twitter data, we use Twitter Streaming API. After collecting data, We implemented event detection system by analyze the frequency of a keyword which contained in a twit in a particular time and clustering the keywords that describes same event by exploiting keywords' co-occurrence graph. Finally, we evaluates the validity of our method through experiments.

☞ keyword : Event Detection, Social Network, Social Media Contents

1. 서 론

소셜 네트워킹 서비스들이 출시된 후 인기를 끌면서 많은 사용자들이 소셜 네트워킹 서비스를 이용하고 있다. 소셜 네트워킹 서비스들을 통해 사용자들은 자신의 일상생활이나 주변 상황을 개인적인 생각을 추가하거나 묘사를 하여 공유한다. 이러한 데이터들을 수집하여 분석한다면 특정 시간, 지역에서 어떤 사건이 관심을 받고 있는지를 알아낼 수 있기 때문에 매우 유용하게 사용할 수 있을 것이다.

본 논문에서는 트위터를 통해 수집된 데이터 중 지역정보를 가진 데이터들을 이용하여 특정 시간, 지역에 다수의 사용자들이 관심을 가지고 있는 일, 즉 이벤트를 분석하기 위한 방법을 제안한다. 이를 위해 트위터 스트리밍 API를 이용해 데이터를 수집하였고 키워드들의 시간에 따른 빈도수를 패턴을 분석하여 정상적인 패턴과는 다른 패턴을 가지는 키워드를 이벤트로 분류한다. 하나의 이벤트를 다른 키워드들이 설명할 수 있기 때문에 비슷한 의미를 가지는 키워드들을 군집화 하기 위해 co-occurrence 그래프를 사용한다. 이벤트를 감지하는 방법을 제안한 기존 연구들이 사용한 데이터는 정확한 위도와 경도가 표시되어 있어 공간과 시간을 동시에 분석하는 방법을 사용했다. 본 연구에서 이용하는 데이터는 위도와 경도에 대한 정보가 없고 지역구에 대한 정보 (예를 들어 서울시 송파구, 성남시 수정구) 만을 사용할 수 있기에 이벤트에 대한 키워드들을 군집화하려 키워드 co-occurrence 그래프를 이용하였다.

[1] Dept. of Software, Gachon Univ., Seongnam, 461-701, Korea)
* Corresponding author (orjeong@gachon.ac.kr)
[Received 31 October 2016, Reviewed 7 Nvember 2016, Accepted 9 December 2016]
☆ 본 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 기초연구사업지원과 미래창조과학부 및 정보통신기술진흥센터의 ICT/SW창의연구과정지원사업(SW중심대학)의 지원을 받아 수행한 것임. (NRF-2015R1C1A2A01051729, R2215-14-1006)

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 방법과 연관된 관련 연구 및 3장에서는 제안하는 방법의 구조와 각 모듈에 대해 설명한다. 4장에서는 제안하는 방법을 검증하기 위한 실험과 실험 결과를 보여주고 5장에서는 결론 및 추후 계획을 기술한다.

2. 관련 연구

이벤트 감지에 관한 연구는 관심을 받고 있는 주제들을 감지하는 Topic Detection의 일부분으로 진행이 되어 왔다. Topic Detection과는 다르게 이벤트 감지는 새로운 상품의 출시, 재난상황 발생 등과 같이 특정 시간과 장소에서 발생하여 다수의 사람들에게서 관심을 받고 있는 일들을 감지해 내는 것이 목적이다. 이벤트 감지를 하기 위해 데이터를 얻고자 뉴스 데이터를 대상으로 하였고, TF-IDF 방법을 적용하여 각 뉴스 문서를 term vector 또는 bag of words로 표현하였다. Term vector, bag of words로 바뀐 뉴스 문서들을 이용하여 각 키워드의 빈도수가 단기간에 치솟는 것을 이벤트로 간주하여 감지해내는 방법을 feature-pivot 방식이라 한다. 최근 이벤트 감지에는 실시간으로 사용자들이 주변 상황이나 자신의 경험을 게시하는 소셜 네트워크 데이터들을 주로 이용하고 있다. 하지만 뉴스와는 달리 대표적인 소셜 네트워크중 하나인 트위터는 게시글의 문자 수가 140자로 제한 되어 있기 때문에 줄임말, 강조, 이모티콘 등의 약어들을 많이 사용한다. 때문에 전통적인 이벤트 감지 방법을 적용했을 때 성능이 기대에 미치지 못한다. [1-3]은 이벤트를 감지하고 감지된 이벤트에 대한 정보들을 보여줄 수 있는 TwitterMonitor, KeySEE, TwitInfo를 제안하였다. 감지된 이벤트와 이벤트에 관련된 키워드들을 차트, 타임라인 등을 이용하여 보여준다. [4-7]은 클러스터링 기법을 이용하여 이벤트를 감지하며 각각의 클러스터를 Naïve bayes classifier, support vector machine등을 이용하여 이벤트가 맞는지 판별하는 방법을 제안하였다. [8-9]는 단기간에 빈도수가 증가하는 이벤트에 관련된 키워드들을 감지하기 위해 시간-빈도수로 표현된 키워드들을 Discrete Fourier Transform을 이용한 방법을 제안하였다.

본 논문에서는 트위터 API 정책의 변경으로 인해 Tweet의 위도와 경도에 관한 데이터를 수집하지 못하므로 지역구 정보를 이용하며 키워드 co-occurrence graph를 이용하여 같은 이벤트를 설명하는 키워드들을 간단하게 군집화 시키는 방법을 제안한다.

3. 이벤트 감지 시스템

그림 1은 제안하고 있는 시스템 구조이다. 제안하는 시스템은 데이터 필터링 모듈, 키워드 패턴 분석 모듈, 이벤트 감지 모듈로 이루어져 있다. 데이터 필터링 모듈은 트위터 스트리밍 API를 이용하여 데이터를 수집하며, 수집된 데이터 중 지역정보를 가진 데이터를 트윗만을 필터링하여 시간, 트윗 내용 그리고 지역구 정보만을 저장한다. 키워드 패턴 분석 모듈은 시간에 따른 키워드 빈도수를 분석하여 키워드의 정상적인 패턴과는 다른, 키워드들을 골라낸다. 그리고 이벤트 감지 모듈은 키워드 패턴 분석 모듈에서 선정된 이벤트에 관련된 키워드들을 co-occurrence 그래프를 이용하여 비슷한 키워드들끼리 군집화한다.

3.1 데이터 필터링 모듈

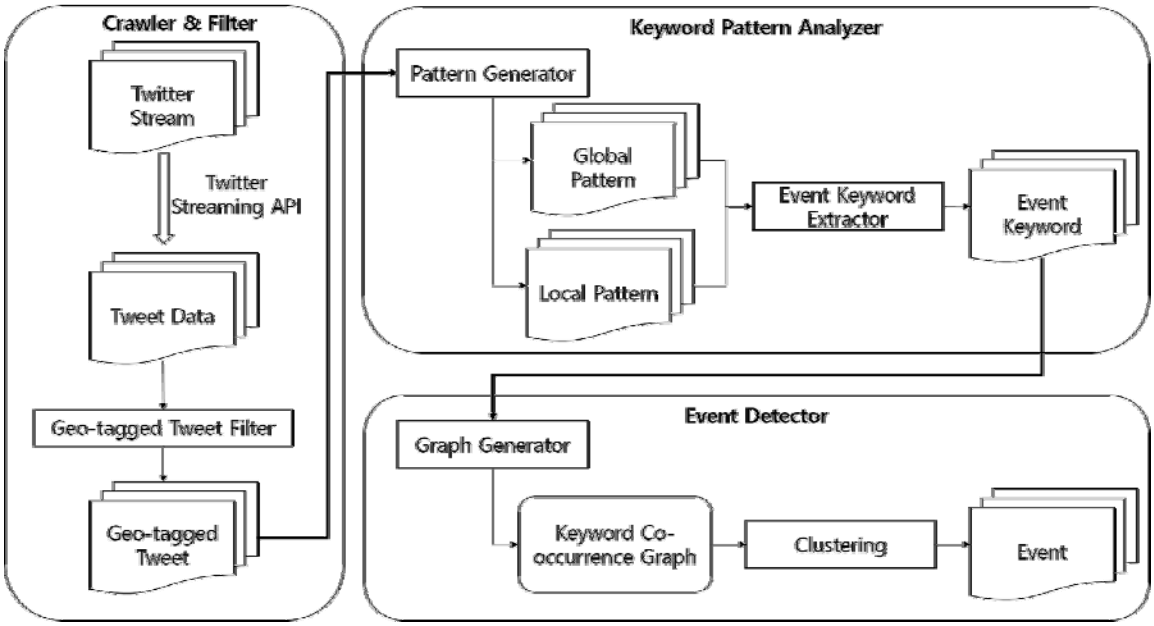
한국 트윗 데이터를 수집하기 위해 트위터 스트리밍 API중 filter API를 이용하였다. 이는 사각형 범위안의 지역에 게시된 트윗의 1%를 임의로 제공하고 있어 계정을 인증 한 뒤 API 요청을 하면 손쉽게 데이터를 수집할 수 있다. 데이터 필터링 모듈은 수집된 데이터 중 지역정보를 가진 데이터를 골라내고 필요한 키워드들만을 골라내는 작업을 한다. 수집된 트윗은 사용자 정보, 트윗 내용, 게시된 날짜 그리고 지역 정보로 나누어져 있지만 트위터에서 수집된 데이터는 지역정보를 가지고 있지 않은 데이터가 대부분이다. 본 논문에서 사용해야 하는 데이터는 지역정보가 필요하기 때문에 위치정보가 있는 데이터만을 골라내었다.

특정 이벤트에 대한 설명이나 이벤트 자체는 명사이기 때문에 명사만을 골라내 시스템에 이용하였다. 명사만을 골라내기 위해 한국어 트윗 처리기인 twitter-korean-text를 이용하였다. Twitter-korean-text는 트위터에서 만든 오픈소스 한국어 처리기이며 트윗의 정규화와 형태소 분석, 스테밍을 지원하고 있다.

3.2 키워드 패턴 분석 모듈

이벤트가 발생했을 때 사용자들은 소셜 네트워킹 서비스를 이용하여 그 이벤트에 대한 설명이나 사용자의 생각 등을 게시할 것이며 이벤트에 관련된 게시글들의 수가 증가할 것이다.

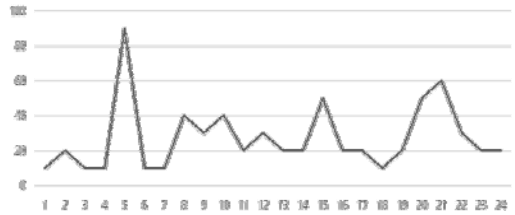
우리는 이를 이용하여 이벤트를 감지하기 위해 시간-빈도수 패턴을 이용하였다. 그림 2와 3은 하루 동안의 특



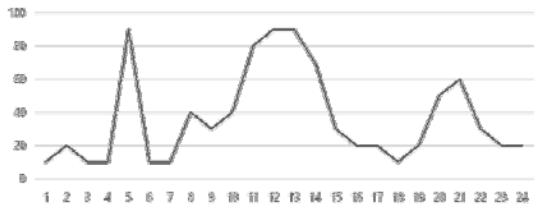
(그림 1) 이벤트 감지 시스템 구조
(Figure 1) Event Detection System Structure

정 키워드 시간-빈도수 패턴의 예이다. X축은 시간, Y축은 빈도수를 나타낸다. 그림 2는 키워드의 정상적인 패턴의 예이고 그림 3은 키워드의 비정상적인 패턴의 예이다. 그림 2와 다르게 그림 3에서는 11시~14시에 키워드의 빈도수가 치솟았다. 이를 감지하기 위해 우리는 특정 키워드의 시간-빈도수 패턴을 하루 단위로 나누어 평균을 내 키워드의 정상적인 패턴으로 사용한다.

키워드의 현재 t시와 t+1시 사이의 빈도수를 f_t, t시와 t+1시 사이의 키워드 빈도수의 평균인 정상적인 패턴을 〖nf〗_t 그리고 〖nf〗_t의 표준편차를 σ_t라 했을 때 〖nf〗_t+σ_t<f_t를 만족하는 키워드를 이벤트로 감지한다.



(그림 2) 키워드의 정상적인 패턴 예
(Figure 2) Example of normal keyword pattern



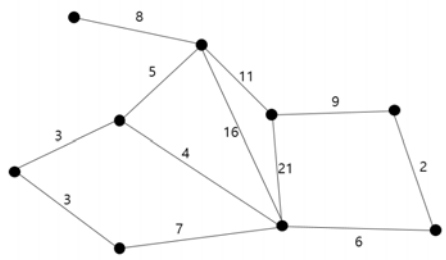
(그림 3) 키워드의 비정상적인 패턴 예
(Figure 3) Example of abnormal keyword pattern

3.3 이벤트 감지 모듈

하나의 이벤트에 대한 설명이나 이벤트 자체가 비슷한 여러 개여 키워드들로 이루어 질 수 있기 때문에 키워드 co-occurrence 그래프를 이용하여 비슷한 키워드들을 군집화 한다. 그림 4는 키워드 co-occurrence 그래프의 예시이다. 각 점은 키워드를 의미하며 점과 점 사이의 선은 두 키워드가 동시에 출현한 횟수로 표시된다.

군집화를 하기 위해 DB-SCAN 알고리즘을 응용하여 사용하였다. 키워드 co-occurrence 그래프에는 좌표값이 존재하지 않기 때문에 클러스터링에 이용할 node간의 거리는 같이 출몰한 빈도를 이용하였다. 그림 4 에서 각 키

워드가 동시에 출몰한 빈도수를 가중치로 갖는 키워드 co-occurrence 그래프에서 각 node간의 거리 d는 같이 출몰하는 빈도수의 역수 1/f로 정의하며, 거리 d가 임계치 D를 넘지 않는 키워드의 수가 N을 넘는 키워드를 중심으로 군집화를 한다.



(그림 4) 키워드 co-occurrence 그래프
(Figure 4) Keyword co-occurrence graph

4. 구현 및 실험

본 논문에서 제안하는 시스템이 이벤트를 잘 감지하는지 검증하기 위해 구현하고 실험하였다. 데이터 셋으로는 트위터 스트리밍 API를 통해 수집한 2016년 8월 한국 데이터를 이용하였다. 수집한 트윗은 총 649,290개이다.

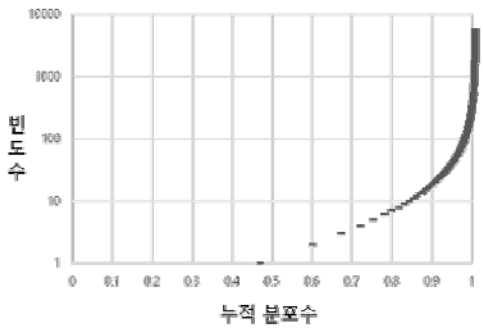
4.1 실험 방법

본 논문에서 제안하는 시스템을 검증하기 위해 기존 연구인 [10]과 비교한다. [10]은 위도와 경도를 0.5씩 나눠 grid화하여 하나의 grid안에서 키워드의 빈도수가 치솟는 정도를 사용하여 이벤트를 감지하였다. 우리는 이 연구에서 사용한 grid를 각 지역구로 바꾸어 실험에 사용하였다. 그림 5는 데이터 필터링 모듈을 거쳐서 나온 모든 트윗들에 포함된 키워드들의 빈도수를 누적 분포수 그래프로 나타낸 것이다.

이는 Twitter Streaming API를 통해 가져올 수 있는 데이터의 문제점을 보여주고 있다. 8월 한달동안 키워드의 출몰 빈도수가 1000을 넘기는 것의 비율이 0.15%이며 100을 넘기는 키워드의 비율은 2.6%에 불과하다. 따라서 오랜 시간 데이터를 축적하여 사용하지 않는 경우에 사용하기에 부적절하다. 본 연구에서 사용한 데이터는 오랜 시간 축적하지 않았지만 출몰 빈도수가 많은 키워드들을 감지하는데 초점을 맞추어 실험을 진행하였다.

4.2 실험 결과

우리는 제안하는 시스템을 검증하기 위해 8월 데이터를 이용하였다. 8월 데이터를 이용하여 정상적인 키워드 패턴들을 만들었으며 이를 이용하여 9월에 발생한 이벤트를 감지해 냈다.



(그림 5) 키워드 출몰 빈도수의 누적분포그래프
(Figure 5) Keyword occurrence CDF graph

표 1은 제안하는 방법과 비교실험인 [10]의 9월 이벤트 감지 결과이다. 9월 7일 아이폰의 새 제품의 출시를 비롯하여 9월에 발생한 규모 5.8의 지진이 이벤트로 감지되었다. 표 2은 통계청에 보도된 지진 기록이다. 지진이 발생한 지역과 시간을 나타내고 있다.

(표 1) 9월에 감지된 이벤트
(Table 1) Detected events in September

날짜	제안하는 방법		비교실험[10]	
	이벤트	지역	이벤트	지역
9/7	아이폰	대구, 서울, 부산		
9/12	지진, 경주, 문자	경주, 울산, 부산	지진	경주, 부산
9/13	지진, 경주	서울, 부산	지진	서울, 부산
9/19	지진	김해. 부산, 울산	지진	부산
9/20	지진	서울, 부산, 경주		
9/21	지진	서울, 포항, 부산		

(표 2) 통계청 지진 기록
(Table 2) Earthquake records of Korea

날짜	규모	발생지역
9/12	2.0 ~ 6.0	경주, 울산
9/13	2.0 ~ 4.0	경주
9/14	2.0 ~ 4.0	경주, 전라남도 완도
9/15	2.0 ~ 3.0	경주
9/16	2.0 ~ 3.0	경주
9/17	2.0 ~ 3.0	경주
9/18	2.0 ~ 3.0	경주
9/19	2.0 ~ 5.0	경주
9/20	2.0 ~ 3.0	경주, 강원도 양구
9/21	2.0 ~ 5.0	경주
9/22	2.0 ~ 3.0	경주
9/23	2.0 ~ 3.0	경주
9/24	2.0 ~ 3.0	경주
9/28	2.0 ~ 4.0	경주
9/30	2.0 ~ 3.0	경주

표 3은 지진 이벤트 감지에 대한 실험 결과이다. 비교 실험의 결과가 하루 단위이기 때문에 제안하는 방법의 결과를 하루 단위로 바꾸어 비교하였다.

(표 3) 실험 결과
(Table 3) Experiment Result

규모	Precision	Recall	F-score
제안하는 방법	1.0	0.77	0.87
비교 실험	1.0	0.72	0.84

본 논문에서는 지진에 관한 정보를 감지하는 것이 아니므로 규모에 대한 수치를 감지해내지 않는다. 따라서 규모에 상관 없이 감지된 지진 이벤트를 이용하여 제안하는 방법의 성능을 평가하였다. Precision은 제안하는 방법과 비교 실험 모두 1.0으로 같았으나 제안하는 방법의 Recall과 F-score가 각각 0.05, 0.03 향상되어 0.77, 0.87의 결과를 보이고 있다.

5. 결 론

본 논문에서는 지역정보를 가진 트윗 데이터들을 이용하여 특정 시간, 지역에 발생하는 이벤트를 감지하기 위한 방법을 제안했다. 이벤트를 감지하기 위해 키워드들의 시간에 따른 빈도수를 패턴을 분석하여 정상적인 패턴과는 다른 패턴을 가지는 키워드를 이벤트로 분류하여 감지했다. 또한 같은 이벤트를 설명하는 여러 단어들을 큰 범주의 키워드로 군집화 하기 위해 co-occurrence 그래프를 사용하였다. 제안한 방법을 검증하기 위해 통계청에 보도된 우리나라에서 최근에 발생한 지진과 그에 따른 여진기록을 이용하였다. 실험 결과 통계청에 보도된 지진이 잘 감지되었으나 다른 지역에서 일어난 지진이 감지되는 문제가 있다. 추후 이를 해결하기 위해 트윗 내용을 이용하여 지역을 예측하는 방법을 도입하면 결과가 더 향상될 것으로 기대된다.

참 고 문 헌 (Reference)

[1] Mathioudakis, Michael, and Nick Koudas. "Twittermonitor: trend detection over the twitter stream." Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010. https://doi.org/10.1145/1807167.1807306

[2] Lee, Pei, Laks VS Lakshmanan, and Evangelos Milios. "Keysee: Supporting keyword search on evolving events in social streams." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013. https://doi.org/10.1145/2487575.2487711

[3] Marcus, Adam, et al. "Twitinfo: aggregating and visualizing microblogs for event exploration." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2011. http://dl.acm.org/citation.cfm?doid=1978942.1978975

[4] Sankaranarayanan, Jagan, et al. "Twitterstand: news in tweets." Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, 2009. https://doi.org/10.1145/1653771.1653781

[5] Phuvipadawat, Swit, and Tsuyoshi Murata. "Breaking news detection and tracking in Twitter." Web Intelligence and

Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 3. IEEE, 2010. https://doi.org/10.1109/WI-IAT.2010.205

[6] Petrović, Saša, Miles Osborne, and Victor Lavrenko. "Streaming first story detection with application to twitter." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. http://dl.acm.org/citation.cfm?id=1858020&CFID=878446591&CFTOKEN=12113260

[7] Becker, Hila, Mor Naaman, and Luis Gravano. "Selecting Quality Twitter Content for Events." ICWSM 11 (2011). http://www.cs.columbia.edu/~gravano/Papers/2011/icwsm11-b.pdf

[8] Weng, Jianshu, and Bu-Sung Lee. "Event Detection in Twitter." ICWSM 11 (2011): 401-408. http://www.hpl.hp.com/techreports/2011/HPL-2011-98.pdf

[9] He, Qi, Kuiyu Chang, and Ee-Peng Lim. "Analyzing feature trajectories for event detection." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007. https://doi.org/10.1145/1277741.1277779

[10] Kaneko, Takamu, and Keiji Yanai. "Visual Event Mining from the Twitter Stream." Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016. https://doi.org/10.1145/2872518.2889418

◐ 저 자 소 개 ◑

박 태 수 (Tae Soo Park)
2015년 가천대학교 소프트웨어 설계·경영학과 졸업(학사)
2015년~현재 가천대학교 일반대학원 소프트웨어 설계·경영학과 (석사과정)
관심분야: 데이터 마이닝, 소셜 네트워크

정 옥 란 (Ok-Ran Jeong)
2005년 이화여자대학교 컴퓨터공학과 (공학박사)
2005년~2006년 서울대학교 컴퓨터공학부 (박사후 연구원)
2007년 Univ. of Illinois of Urbana Champaign (visiting scholar)
2008년~2009년 성균관대학교 정보통신학부 (연구교수)
2009년~2015년 가천대학교 소프트웨어 설계·경영학과 (조교수)
2015년~현재 가천대학교 소프트웨어학과 (부교수)
관심분야 : 웹 마이닝, 정보검색, 추천 시스템, 소셜 컴퓨팅
E-mail : orjeong@gachon.ac.kr