한국지능시스템학회 논문지 2011, Vol. 21, No. 5, pp. 564-568
http://dx.doi.org/10.5391/JKIIS.2011.21.5.564

마이크로블로그를 통한 그래프 기반의 토픽 추출에 관한 연구

A Study on Graph-based Topic Extraction from Microblogs

최돈정*・이성우*・김재광*・이지형*†
Don-Jung Choi*, Sung-Woo Lee*, Jae-Kwang Kim* and Jee-Hyong Lee*†

* 성균관대학교 전자전기컴퓨터공학과

요 약

오늘날 마이크로블로그는 스마트폰의 보급과 더불어 대중적인 정보전달 방식의 하나로 자리 잡고 있으며, 기존의 정보매체에 비해 사용자들의 관심사 변화를 보다 빠르게 반영하는 특징을 지닌다. 특히 다수 사용자의 관심을 끌고 있는 토픽의 경우, 다양한 정보 출처로부터 풍부한 정보를 제공할 수 있는 잠재력을 보유하고 있기도 하다. 그럼에도 불구하고 높은 비율로 존재하는 노이즈 등으로 인해 마이크로블로그로부터 유용한 정보를 획득하기란 쉽지 않은 문제로 남아있다. 지금까지 특정 문서로부터 주제를 효율적으로 추출, 추적하는 다양한 방법이 제안되었으나, 마이크로블로그와 같은 단문의 문서가 대량으로 생산되는 경우에 활용하기에는 미흡한 점이 있었다. 본 논문에서는 특정 주제어가 주어졌을 때, 키워드 그래프를 구성함으로써 그에 대한 사용자들의 관심사가 어떻게 변화하는지를 효과적으로 파악하는 방법을 제안한다. 제안 방법은 크게 마이크로블로그 내에서의 단어 동시출현빈도를 이용하여 단어간 키워드 그래프를 생성하는 과정과, 네트워크 분할 기법을 이용하여 그래프를 적절히 분할함으로써 사용자의 관심사 별로 나누는 과정을 포함한다. 선별된 주제어에 대해 제안된 방법을 적용해 봄으로서 적은 비용으로 효과적인 주제 발견 및 분할이 가능함을 확인하였다.

키워드 : 마이크로블로그, 트위터, 키워드 그래프, 네트워크 분할.

Abstract

Microblogs became popular information delivery ways due to the spread of smart phones. They have the characteristic of reflecting the interests of users more quickly than other medium. Particularly, in case of the subject which attracts many users, microblogs can supply rich information originated from various information sources. Nevertheless, it has been considered as a hard problem to obtain useful information from microblogs because too much noises are in them. So far, various methods are proposed to extract and track some subjects from particular documents, yet these methods do not work effectively in case of microblogs which consist of short phrases. In this paper, we propose a graph-based topic extraction and partitioning method to understand interests of users about a certain keyword. The proposed method contains the process of generating a keyword graph using the co-occurrences of terms in the microblogs, and the process of splitting the graph by using a network partitioning method. When we applied the proposed method on some keywords. our method shows good performance for finding a topic about the keyword and partitioning the topic into sub-topics.

Key Words : Microblogs, Twitter, Keyword graph, Network partitioning.

1. 서 론

스마트폰의 폭발적인 보급과 더불어, 마이크로블로그서비스가 대중적인 정보전달 방식의 하나로 자리 잡고 있다. 마이크로블로그의 대표적인 예라 할 수 있는 트위터의 경우 2011년 기준으로 이미 2억명이 넘는 사용자가 이용하고 있으며, 하루 평균 2억 개의 마이크로블로그가 생성되고 16억 개의 검색어를 수용하고 있다 [1].

마이크로블로그는 짧은 글자 수 제한과 모바일에 최적화된 쉽고 빠른 정보 전송을 무기로 시간과 공간의 제약 없이 누구나 자신의 의견을 간단한 문장 형태로 남길 수 있다. 또한 대부분의 마이크로블로그는 정보의 개방성에 기반을 두고 있어 누구나 열람할 수 있다는 특징을 가진다. 많은 사용자들이 마이크로블로그를 통해 특정 상품, 인물, 사건 등에 대한 자신의 의견을 피력하기도 하며, 유용한 정보를 공유, 검색하거나 사적인 대화를 나누기도 한

접수일자 : 2011년 3월 19일
완료일자 : 2011년 10월 10일
†Corresponding Author : jhlee@ece.skku.ac.kr
본 논문은 본 학회 2011년도 춘계학술대회에서 선정된 우수논문입니다.
감사의 글 : 본 연구는 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업의 연구 결과입니다.

다 [2].

마이크로블로그를 통해 전달되는 정보는 다양한 형태로 대중의 관심사를 파악하는데 사용될 수 있다. 이미 많은 기업에서 자사 제품에 대한 세간의 평가를 확인하기 위해 마이크로블로그로부터 정보를 수집하고 있으며, 선거와 관련된 후보에 대한 대중의 관심사 파악, 특정 속보의 발빠른 확인 등에도 마이크로블로그가 유용하게 사용된다. 그러나 기하급수적으로 늘어나고 있는 마이크로블로그의 생성량은, 역설적이게도 마이크로블로그로부터 유용한 정보를 획득하는 것을 더욱 힘들게 하고 있다. 앞서 예로 든 트위터의 경우 자체 검색 기능이 존재하지만, 특정 검색어에 대해 트윗(마이크로블로그에 해당)을 시간대별로 보여주는 것에 그치고 있어 정보의 파악이 힘든 문제가 있으며, 많은 트위터용 검색 엔진이 생겨나고 있는 것도 이러한 이유 때문이다.

지금까지 마이크로블로그로부터 유용한 정보를 얻기 위한 다양한 시도가 이루어졌다. [3]에서는 실시간으로 수집되는 마이크로블로그의 컨텐츠와 사용자의 RSS 피드 간 유사도를 비교함으로써 사용자에게 뉴스를 추천해주는 기법을 제안하고 있으며, [4]는 위키피디아 정보를 이용하여 마이크로블로그 내의 개체를 분류해놓음으로써 사용자들이 관심을 가지는 주제를 찾는 방법을 다루고 있다. 그러나 이러한 시도는 특정 주제에 대해 미묘하게 나뉘어질 수 있는 사용자들의 관심사에 대한 고려가 부족하며, 마이크로블로그 외적인 정보를 필요로 하는 문제점이 있다.

그 외에도 특정 문서로부터 주제를 추출하는 다양한 방법들이 제안되었다 [5][6]. 위키피디아를 이용하여 단어 간 의미적인 관련 정도를 계산하고, 이를 그래프로 변환하여 주제를 추출하거나 [5], 문서에서의 단어 빈도 분포를 가우시간 혼합 모델로 생각하여 EM 알고리즘을 적용하여 주제를 찾는 방법도 연구되었다 [6]. 이와 같은 연구는 다량의 문장으로 구성된 문서에 초점을 맞추고 있기 때문에 짧은 문장으로 이루어진 마이크로블로그에는 적합하지 않으며, 주제를 추출하는데 비용 또한 많이 소요된다는 문제가 있다.

본 논문에서는 소수 개의 짧은 문장으로 구성되며, 특정 주제에 대해 대중의 관심사가 반복적인 용어의 사용으로 나타나는 마이크로블로그의 특성을 이용하여 효율적으로 관심사를 분리해내는 방법을 제안한다. 또한 실험을 통해, 임의 선별된 주제에 대해 대중의 관심사 변화를 효율적으로 추적할 수 있음을 보인다.

2. 그래프 기반의 관심사 파악

제안하는 방법은 크게 2단계로 구분된다. 첫 번째 단계는 특정 주제어를 포함하는 마이크로블로그를 찾아 등장하는 단어들에 대해 키워드 그래프를 생성하는 단계이며, 두 번째 단계는 생성된 키워드 그래프를 바탕으로 네트워크 분할 기법을 적용하여 세부 주제들로 나누는 과정이다. 세부 과정은 다음과 같다.

2.1 키워드 그래프의 생성

특정 주제어에 대해 대중의 관심사를 파악하기 위해 우리는 사용자들이 작성한 마이크로블로그를 이용하여 키워드 그래프를 생성한다. 우선 마이크로블로그 내에 쓰여진 각 단어들을 대상으로 키워드 자체와 불용어를 제거한 후 하나의 노드(node)에 대응시킨다. 다음으로 동일한 마이크로블로그 내에 사용된 단어들 간에는 에지(edge)를 구성하여 가중치를 증가시킨다. 가령 “Steve Jobs″라는 주제어에 대해 “Steve Jobs died because of pancreatic cancer. I'll remember him forever.″ 라는 마이크로블로그가 있다면 ‘die, pancreatic, cancer, I, remember, him, forever’가 그래프의 노드를 구성하게 된다.

이는 임의의 사용자가 특정 주제어를 포함하는 글을 게시할 때 그 사용자의 관점이 포함되어 있을 것이라는 가정에 기초한다. 즉, 앞선 예시에서 스티브잡스에 대한 사용자의 관심은 췌장암(pancreatic cancer)으로 인한 사망 사실과 그를 영원히 기억하겠다는 것이다. 만약 동일한 시기에 많은 마이크로블로그에서 스티브잡스를 애도하는 글이 생겨나고 있다면, 그를 애도하는 몇몇 단어들이 많은 동시 발생빈도를 나타낼 수 있으며, 그러한 단어들 사이에는 높은 가중치의 에지가 형성된다.

다수 개의 마이크로블로그에 등장한 단어를 노드로, 단어 간 연관성을 가중치를 가지는 에지로 대응시키는 과정은 그림 1과 같다.

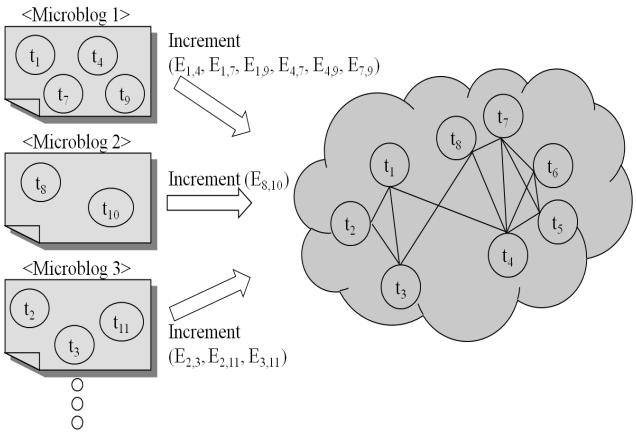

그림 1. 키워드 그래프 생성.
Fig. 1. Generating Keywords Graph.

2.2 네트워크 분할 기법을 이용한 세부 토픽 발견

전체 그래프를 부분 그래프로 분할하는 다양한 방법이 연구되어 있다. 우리는 2.1절에서 생성한 키워드 그래프를 최적의 상태로 자연분할 함으로써 특정 주제어를 포괄하는 하나의 그래프를, 세부 주제를 표현하는 부분 그래프로 분할하고자 한다. 이 과정을 통해 복잡하고 다양한 대중의 관심사를 받고 있는 특정 주제에 대해 관심사 별로 마이크로블로그를 요약하는 것이 가능해진다.

본 논문에서는 M. E. J. Newman과 M. Girvan이 제안한 네트워크 분할 방법[7]을 이용해 키워드 그래프를 세부 주제별로 클러스터링한다. [7]에서 제안하는 방법은 노드 간의 거리를 이용해 전체 네트워크를 고밀도의 하위 네트워크로 자연 분할하는데 유용하다. 이들 연구의 핵심은 반복적으로 노드 사이의 매개성(betweenness)을 계산하고, 가장 큰 매개성을 가지는 에지를 제거해 나감으로써 네트워크의 모듈성(modularity)이 가장 큰 상태가 될 때까지 네트워크를 분할해 나가는 것이다. 여기서 에지의 매개성은 임의의 모든 노드로부터 다른 모든 노드로 최단거리

이동을 한다고 할 때, 특정 에지가 몇 번이나 이용되는지를 나타낸다. 그래프의 중심부에 위치하면서 소수 개의 노드와 인접한 에지일수록 매개성이 커지게 되며, 이러한 에지를 우선 제거함으로써 전체 그래프를 부분 그래프로 분할할 수 있다. [7]에서 제안하는 최단 거리 매개성 계산을 따를 경우, 임의의 시작점에서 최단 거리 경로가 하나가 아닐 때 n개의 노드와 m개의 에지에 대해 O(m2n) 또는 O(n3)(희소 그래프의 경우)의 시간 복잡도로 네트워크를 분할할 수 있다.

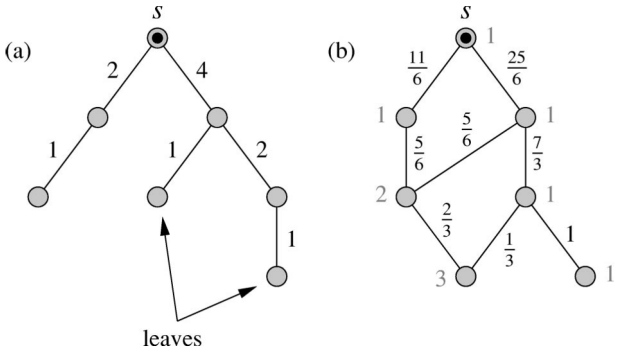

그림 2. 최단 거리 매개성의 계산
Fig. 2. Calculation of shortest-path betweenness

```
cal_betweenness (s)
input : source vertex
output : all betweeness scores

//assign (dist(=distance), weight) to all vertices
s.dist = 0, s.weight = 1
for (all adjacent vertices i at vertex s)
  i.dist = s.dist+1, i.weight = s.weight
end for
while (there exist direct lower vertices j for all i)
  if (j.dist is not assigned)
    j.dist = i.dist+1, j.weight = i.weight
  else if (j.dist is assigned and j.dist = i.dist+1)
    j.weight = j.weight + i.weight
  else
    do nothing
  end if
  vertices i = vertices j
end while

//assign betweeness scores to edges e_{x,y}
find all leaf vertices t
for (all direct upper vertices i at t)
  score (e_{i,t}) = i.weight / t.weight
end for
while (vertices j(direct upper vertices at i) do not
       reach to vertex s)
  e_{i,j} = 1 + ∑(all e immediately below it) ×
       (i.weight / j.weight)
  edges i = edges j
end while
```

그림 3. 매개성 점수 계산 알고리즘
Fig. 3. Algorithm to calculate betweenness score

그림 2는 [7]에서 제안한 최단 거리 매개성의 계산에 있어서, 두 가지 형태의 서로 다른 그래프를 보여주고 있다. 그림 2의 (a)는 임의의 시작점에서 다른 모든 노드로 향하는 최단 거리 경로가 단 하나만 존재하는 경우를 나타내며, 그림 2의 (b)는 임의의 시작점에서 다른 노드로의 최단 거리 경로가 여러 개인 경우를 보여준다. (a)의 경우보다 간단한 알고리즘으로 m개의 에지와 n개의 노드에 대해 O(mn)의 시간 복잡도가 소요됨이 증명되었으나, 우리가 제안하는 방법은 (a) 또는 (b)의 형태를 확정할 수 없기에 보다 시간 복잡도가 높은 (b)의 형태를 가정한다. 따라서 최단 거리 매개성 경로를 이용한 네트워크 분할 알고리즘은 그림 3과 같다.

3. 실험 및 분석

3.1 데이터 수집 및 분석

트위터 Open API를 이용해 2010년 3월 10일부터 3월 16일까지 일주일간 'Japan' 검색어를 포함하는 마이크로블로그를 수집하였다. 각 날짜별 2천 개의 마이크로블로그를 무작위로 선택하였으며, 총 1만 4천 개의 마이크로블로그가 분석에 사용되었다. 수집된 모든 마이크로블로그는 자연히 검색어 'Japan'을 포함하고 있다. 우리가 관심을 가지는 것은 'Japan'에 대해 사용자들의 관심이 반영된 정보를 찾는 것이므로, 전처리 과정을 통해 모든 마이크로블로그로부터 'Japan'이라는 단어는 제거하였다. 또한 Java 환경에서 Lucene[1] 라이브러리를 활용하여 불용어 처리 및 형태소 분석 과정을 전처리 단계에서 수행하였다.

2.1절에서 제안한 방법에 의해 토픽과 연관된 단어들의 쌍을 누적한 결과는 그림 4와 같다. 일주일간의 각 날짜별 상위 500개의 단어 쌍에 대해, 단어 간 동시 발생 빈도는 긴 꼬리 형태를 띠고 있으며 많은 동시 발생 빈도를 가지는 단어 쌍의 집합이 토픽 'Japan'과 관련한 관심사를 반영함을 유추할 수 있다.

1) http://lucene.apache.org/

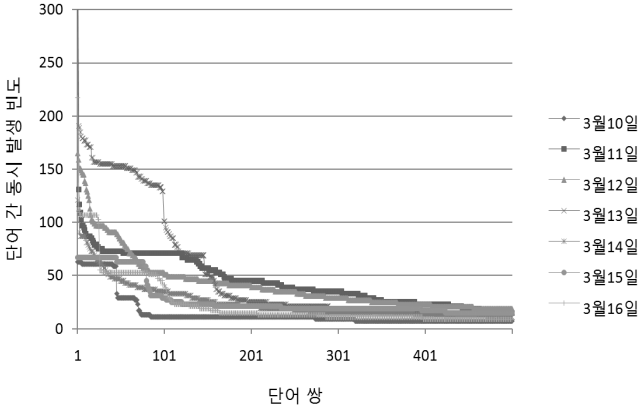

그림 4. 단어 간 동시 발생 빈도.
Fig. 4. The frequencies of co-occurrence between different two terms.

3.2 네트워크 분할을 이용한 세부 주제 파악

단어 간 동시 발생 빈도를 이용해 구성된 키워드 그래프는 2.2절에 제시된 알고리즘을 이용하여 모든 에지의 매개성 점수를 계산하고 이를 바탕으로 가장 큰 매개성을 가지는 에지부터 하나씩 지워나간다.

이 과정은 그래프가 가장 최적의 상태로 분할될 때까지 계속되며, 최적의 상태를 판단하는 기준은 모듈성(modularity) 척도로 판단하였다. k개의 그룹으로 분할되는 그래프를 생각할 때, 에지의 가중치를 고려하는 무향 그래프는 k×k 대칭행렬 e로 나타낼 수 있다. 또한 정규화 과정을 거쳐 행렬의 각 요소 eij는 0과 1사이의 값을 갖도록 조절할 수 있다. 행렬의 트레이스 $\mathrm{Tr}\ e = \sum_i e_{ii}$이고, ai를 e행렬 i번째 행에 있는 모든 요소의 합이라 할 때 모듈성 점수 Q는 식 (1)과 같이 정의된다 [7].

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \|e^2\| \quad (1)$$

수집된 데이터를 위에 제시된 방법으로 분할한 결과는 표 1과 같다. 노이즈 제거 및 이해의 편의를 위해 단어 쌍의 동시 발생 빈도가 30 이상인 경우만을 대상으로 분할하였다. 또한 한정된 공간으로 인해 3월 10일에서 3월 12일까지 3일간 상위 3개의 클러스터만을 표 1에 나타내었다. 표의 이해를 위해 실험을 진행한 기간 동안 일본에서는 큰 지진으로 인해 많은 인명과 재산 피해가 발생하였음을 알린다. 2011년 3월 11일 지진 발생을 전후하여 검색어 'Japan'과 연관된 단어 쌍의 분포가 급격히 변함을 확인할 수 있다.

표 1. 네트워크 분할을 이용한 토픽 연관 단어 분류.
Table 1. Clustering topic-related terms using the network division method.

	클러스터 C1	클러스터 C2	클러스터 C3
3월 10일	track radar ass bonus excuse britneyspears	kimono transformation look take selenagomez	good morning
3월 11일	people recovery needs aid guy tomlinson help prayer please difference give	tsunami donation 90999 prayforjapan shackett message redcross person heart	relief ladygaga donate effort monster buy bracelet design
3월 12일	daiichi nuclear emergency agency safety reactor tell meltdown power cnn cooling fukushima plant	90999 red redcross text tsunami prayforjapan donate earthquake help message	ways supportjapan victim quake retweet ryanseacrest bing

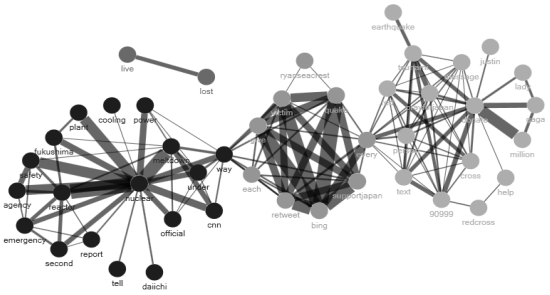

그림 5. 시각화 도구를 이용한 토픽 분류 결과.
Fig. 5. The result of topic clustering using the visualization tool.

그림 5은 3월12일 검색어 'Japan'과 연관된 단어의 분류 결과를 시각화 도구를 이용하여 나타낸 것이다. 단어 쌍의 동시 발생 빈도가 30 이상인 에지(edge)를 대상으로 네트워크 분할 기법을 적용하였다.

4. 결론 및 향후 연구

마이크로블로그는 특정 토픽에 대해 사용자들의 관심변화를 빠르게 반영하는 소셜미디어 중 하나로서, 사회적 관심을 반영하여 토픽에 대한 풍부한 정보를 제공할 수 있다. 그러나 문서로부터 토픽을 추출하기 위한 기존의 연구 방

법은 마이크로블로그로부터 토픽을 추출하는데 적합하지 않았다. 본 논문에서는 마이크로블로그의 특성을 이용하여 적은 비용으로 토픽을 추출하는 방법을 제안하였다. 또한 실험을 통해, 제안된 방법이 적은 비용으로 마이크로블로그로부터 토픽을 추출하고 추적하는데 유효함을 확인할 수 있었다. 그러나 제안된 방법은 큰 관심을 끌지 못하는 토픽에 대해서는 상대적으로 낮은 유효성을 보였으며 스팸 마이크로블로그에 취약한 약점이 있음을 확인하였다. 향후 연구는 이러한 문제점을 개선할 수 있는 방법에 초점을 맞추고자 한다.

참 고 문 헌

[1] http://en.wikipedia.org/wiki/Twitter

[2] A. Java, X, Song, T. Finin and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," *Joint 9th WEBKDD and 1st SNA-KDD Workshop*, 2007.

[3] O. Phelan, K. McCarthy and B. Smyth, "Using Twitter to Recommend Real-Time Topical News," *Proceedings of the 3th ACM conference on Recommender systems*, 2009.

[4] M. Michelson and S. A. Macskassy, "Discovering Users' Topics of Interest on Twitter: A First Look," *Proceedings of the 4th workshop on Analytics for noisy unstructured text data*, 2010.

[5] M. Grineva, M. Grinev and D. Lizorkin, "Extra cting Key Terms From Noisy and Multi-theme Documents," *Proceedings of the 18th Internatio nal Conference on World Wide Web*, 2009.

[6] J. Zeng, C. Wu and W. Wang, "Multi-grain Hierarchical Topic Extraction Algorithm for Text Mining," *Expert Systems with Applications*, Vol. 37(4), pp. 3202-3208, 2010.

[7] M. E. J. Newman, M. Girvan, "Finding and Evaluating Community Structure in Networks," *Journal of Physical Review E*, Vol. 69, 2004.

저 자 소 개

최 돈 정 (Don-Jung Choi)
2010년 아주대 정보 및 컴퓨터공학부 학사
2010년～현재 성균관대 대학원 전자전기컴퓨터공학과 석사과정
관심분야 : 웹, 데이터마이닝, 지능시스템

Phone : +82-31-290-7987
Fax : +82-31-299-4637
E-mail : tenosis@skku.edu

이 성 우 (Sung-Woo Lee)
2009년 성균관대 전자전기컴퓨터공하과 학사
1997년～현재 동 대학원 전자전기컴퓨터공학과 석사과정
관심분야 : 웹, 데이터마이닝, 지능시스템

Phone : +82-31-290-7987
Fax : +82-31-299-4637
E-mail : lsmoney@skku.edu

김 재 광 (Jae-Kwang Kim)
2004년 성균관대 정보통신공학부 학사
2006년 동 대학원 컴퓨터공학과 석사
2006년～현재 동 대학원 전자전기컴퓨터공학과 박사과정
관심분야 : 네트워크 보안, 기계학습

Phone : +82-31-290-7987
Fax : +82-31-299-4637
E-mail : linux@skku.edu

이 지 형 (Jee-Hyong Lee)
1993년 한국과학기술원 전산학과 학사
1995년 한국과학기술원 전산학과 석사
1999년 한국과학기술원 전산학과 박사
2002～현재 성균관대 정보통신공학부 부교수
관심분야 : 퍼지이론, 지능시스템, 기계학습

Phone : +82-31-290-7154
Fax : +82-31-299-4637
E-mail : jhlee@ece.skku.ac.kr