http://dx.doi.org/10.7236/JIIBC.2013.13.5.37
JIIBC 2013-5-5

텍스트 마이닝을 위한 그래프 기반 텍스트 표현 모델의 연구 동향

A Study on Research Trends of Graph-Based Text Representations for Text Mining

장재영*

Jae-Young Chang

요 약 텍스트 마이닝은 비정형화된 텍스트를 분석하여 그 안에 내재된 패턴, 추세, 분포 등의 고급정보들을 추출하는 분야이다. 텍스트 마이닝은 기본적으로 비정형 데이터를 가정하므로 텍스트를 단순화된 모델로 표현하는 것이 필요하다. 현재까지 가장 많이 사용되고 있는 모델은 텍스트를 단순한 단어들의 집합으로 표현한 벡터공간 모델이다. 그러나 최근 들어 단어들의 의미적 관계까지 표현하기 위해 그래프를 이용한 텍스트 표현 모델을 많이 사용하고 있다. 본 논문에서는 텍스트 마이닝을 위한 기존의 연구 중에서 그래프에 기반한 텍스트 표현 모델의 방법들과 그들의 특징들을 기술한다. 또한 그래프 기반 텍스트 마이닝의 향후 발전방향에 대해서도 논한다.

Abstract Text Mining is a research area of retrieving high quality hidden information such as patterns, trends, or distributions through analyzing unformatted text. Basically, since text mining assumes an unstructured text, it needs to be represented as a simple text model for analyzing it. So far, most frequently used model is VSM(Vector Space Model), in which a text is represented as a bag of words. However, recently much researches tried to apply a graph-based text model for representing semantic relationships between words. In this paper, we survey research trends of graph-based text representation models for text mining. Additionally, we also discuss about future models of graph-based text mining.

Key Words : Text Mining, Vector Space Model, Graph-Based Model, Text Representation

Ⅰ. 서 론

텍스트 마이닝(text mining)은 비정형(unstructured) 문서를 대상으로 한 데이터 마이닝(data mining)의 한 분야로서 문서분류(document classification), 군집화(clustering), 인덱싱(indexing), 검색(retrieval), 요약(summarization) 등 문서에 숨겨진 고급 지식들을 탐색하는 분야이다. 특히 최근 들어 빅 데이터(big data) 시대 도래에 따라 대용량 텍스트 데이터 분석기술에 대한 관심이 증대하고 있어, 이 분야의 핵심 기술로서 텍스트 마이닝의 중요성이 더욱 강조되고 있다.

텍스트 마이닝에서는 기본적으로 비정형 데이터를 가정하므로 이들의 분석을 위해서는 정형화된 모델로 추상화하는 작업이 중요하다. 텍스트에 대한 표현 모델로서 지금까지 가장 많이 사용되고 있는 것은 벡터공간 모델(VSM: Vector Space Model)이다[1]. 벡터공간 모델

*정회원, 한성대학교 컴퓨터공학과
접수일자 : 2013년 9월 24일, 수정완료 : 2013년 10월 9일
게재확정일자 : 2013년 10월 11일

Received: 24 September, 2013 / Revised: 10 October, 2013 / Accepted: 11 October, 2013
*Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

에서는 문서에 출현하는 주요 단어와 그들의 가중치(weight)를 벡터 형태로 표현한다. 이와 같이 단순한 표현 방법을 채용한 벡터공간 모델은 현재까지 많은 분야에서 기반 모델로 사용하고 있다. 특히 TF-IDF[2]와 같은 전통적인 검색 기법뿐만 아니라 최신의 검색 기술에서도 꾸준히 채용되고 있다. 벡터공간 모델의 장점은 단순성에 있다. 따라서 문서를 표현하고 처리하는데 있어서 공간과 시간 면에서 많은 비용을 요구하지 않는다. 그러나 표현의 단순성으로 인해 문서 내의 의미적(semantic) 요소나 전후 맥락(context)을 충실히 표현하지 못한다는 단점을 안고 있다.

이러한 문제를 해결하기 위해 2,000년대 이후 그래프 기반 텍스트 마이닝에 대한 연구가 활발히 진행되고 있다. 그래프에 기반을 둔 텍스트 표현 모델에서는 텍스트에 존재하는 단어(term 또는 word), 문장(sentence), 단락(paragraph), 개념(concept) 등의 공기(co-occurrence) 또는 기타 관계(relation) 정보를 활용하여 문서의 특징을 보다 정밀하게 표현할 수 있는 장점이 있다. 따라서 문서에 대한 표현력(expressive power)이 증가하여 텍스트 분석의 정확도를 높일 수 있다. 하지만 반대로 벡터공간 모델에 비해 계산량이 많아지고 많은 자원이 소모되는 단점을 안고 있다. 이러한 문제점들은 최근의 비약적인 하드웨어의 발전으로 인해 점점 해소되고 있는 실정이다. 현재까지의 그래프 기반 텍스트 마이닝의 연구를 보면 복잡한 모델부터 단순한 모델까지 다양한 형태의 그래프를 정의하고 있다. 또한 응용분야 또는 연구 목적에 따라서도 모두 다른 모델을 사용한다.

본 논문에서는 다양한 연구에서 제안된 그래프 기반 텍스트 마이닝의 연구 동향을 분석한다. 우선 기본 모델인 벡터공간 모델을 살펴보고, 지금까지 제안된 그래프 기반 텍스트 모델들의 종류를 특성에 따라 분류한다. 또한 그래프 기반 텍스트 마이닝에서 제안되었던 대표적인 알고리즘이나 표현 모델에 대해 살펴보고, 국내의 관련연구 동향에 대해서도 간략히 정리한다. 마지막으로 이들의 장단점을 바탕으로 향후의 그래프 기반 텍스트 모델의 발전 방향 및 전망에 대해 논한다.

II. 벡터공간 모델

벡터공간 모델은 문서를 다차원 유클리디언(Euclidean) 공간의 벡터로 표현한 단순화된 모델이다. 즉, 문서 d가 주어졌을 때, 벡터공간 모델은 d의 특징(feature)을 다음과 같이 단어들의 가중치 집합으로 표현한다.

$$d = \{t_1 : w_1, \ldots, t_n : w_n\} \quad (1)$$

여기서 $t_i\,(i=1,\ldots n)$ 는 단어를 나타내며, $w_i\,(i=1,\ldots n)$ 는 t_i의 가중치를 나타낸다. 이와 같이 벡터공간 모델에서는 단어를 하나의 차원으로 표현하고, 이를 이용하여 문서를 n차원 공간의 하나의 점(point)로 표현한다. 지금까지 가중치 w_i를 계산하는 많은 방법들이 제안되어 왔는데, 가장 잘 알려진 하는 방법이 TF-IDF[2]이다. TF-IDF는 문서 d에 나타난 단어 t_i의 가중치를 다음과 같은 수식으로 계산한다.

$$w_i = TF(d,t) * IDF(t) \quad (2)$$

여기서 $TF(d,t)$는 단어 t가 문서 d에 출현한 횟수를 나타내며, $IDF(t)$는 전체 문서집합 D 중에서 단어 t가 출현한 문서 집합 D_t의 비율을 역으로 계산한 것으로 다음과 같이 계산된다.

$$IDF(t) = \log \frac{1+|D|}{|D_t|} \quad (3)$$

이와 같이 벡터공간 모델은 비정형적인 텍스트 문서를 단순하고 정형화된 모델로 표현함으로써 기존의 데이터 마이닝에서 사용되었던 다양한 알고리즘들을 수정 없이 그대로 적용할 수 있다. 이러한 장점으로 인해 벡터공간 모델은 현재까지도 많은 연구에서 활용되고 있다. 그러나 이 모델은 표현의 단순성으로 인해 다음과 같은 문제점을 안고 있다[38].

- 개념적으로 서로 유사한 문서지만 다른 용어를 사용하였다면 이들에 대한 유사성(similarity)을 계산할 수 없다.
- 문서의 의미나 구조 등을 표현할 수 없다.
- 단어들이 서로 독립적이므로 단어 간의 출현 순서나 기타 관련성을 표현할 수 없다.

이러한 문제점들을 해결하기 위해 그동안 다양한 연

구가 진행되어 왔지만, 현재까지 그래프를 이용한 텍스트 표현 모델이 가장 대표적인 해결 방법으로 인식되고 있다.

Ⅲ. 그래프 기반 텍스트 모델의 종류

그래프 기반 텍스트 표현 모델을 적용한 연구에서는 그 목적이 기존의 텍스트 마이닝과 동일하다. 구체적으로, 벡터공간 모델을 이용한 텍스트 마이닝에서는 비정형 텍스트 집합에 대해서 다음과 같은 목적을 위한 방법론들을 제안하고 있다.

- 문서 분류(document classification)[3, 4, 5, 6, 7, 8]
- 문서 군집화(document clustering)[9, 10, 11]
- 문서 요약(document summarization)[12, 13, 14, 15]
- 키워드 추출(keyword extraction)[16, 17]
- 인덱싱(indexing)[18]
- 검색(search)[19]
- 노벨티 탐색(novelty detection)[20]
- 오피니언 마이닝(opinion mining)[21, 24]

마찬가지로 본 논문에서 정리하고 하는 그래프 기반 텍스트 마이닝 연구에서도 이와 동일한 목적을 위해 수행되었으며, 단지 그래프를 이용하여 문제 해결을 시도했다는 점에서만 차이가 있다.

그래프 기반 텍스트 표현 모델에서의 이슈는 어떠한 형태의 그래프를 정의할 것인가와 그래프에 어떠한 내용을 담을 것을 것인가로 나눌 수 있다. 본 장에서는 각각의 이슈에 따라 기존의 연구 동향을 정리한다.

1. 그래프 구조(format)에 따른 분류

문서 또는 문서집합으로부터 도출되는 그래프 G는 다음과 같이 간단히 표현될 수 있다.

$$G = \{V, E\} \quad (4)$$

여기서 V는 노드(node)들의 집합이며 E는 노드 간을 연결하는 간선(edge)의 집합이다. 이러한 구조에 대해서 V와 E의 변화에 따라 다양한 형태의 그래프 정의가 가능하다. 본 논문에서는 이들을 노드의 표현 방식과 간선의 표현 방식에 따라 세부적으로 분류한다.

가. 노드의 표현방식

노드는 텍스트의 세부 요소들을 정의할 때 이용된다. 텍스트의 세부 요소로는 단어, 문장, 문단, 문서 등이 있으며, 의미적 요소인 개념도 포함된다. 이러한 요소들에 대해서 그래프가 모두 같은 종류의 요소만으로 노드를 표현하는지 아니면 두 개 이상의 요소로 노드를 표현하는지에 따라 동종(homogeneous) 또는 이종(heterogeneous) 표현 방식으로 나눌 수 있다. 또한 노드에 가중치를 부여할 것인지에 따라 weighted와 unweighted로 구분할 수 있다.

(1) 동종 표현 vs. 이종 표현

동종 표현에서 가장 흔하게 나타나는 방식이 그림 1과 같이 단어를 노드로 표현하는 것이다[3, 4, 5, 7, 9, 12, 13, 19, 20, 22, 23]. 이 방식에서는 많은 경우 단어 간의 공기 정보로 그래프로 표현한다. 공기 정보란 단어들이 동시에 출현하는 것을 나타내는 것으로 하나의 문서나 문장, 또는 n-size 윈도우(window)내에 두 단어가 동시에 나타나면 이를 간선을 연결하는 형태를 말한다. 이 밖에도 단어 간의 문법적 연관성이나 의미적(semantic) 유사성에 따라 그래프로 표현하기도 한다[13, 19, 24]. 이 방식은 단순한 형태로 인해 구축 및 분석에 필요한 계산비용이 적게 소요되며, 기존의 벡터공간 모델에서 사용되었던 여러 가지 알고리즘들을 그대로 사용할 수 있다는 장점이 있다. 이외에도 문장, 문단, 개념 등을 동종 표현 모델로 사용한 연구도 있다[14, 15, 17, 25, 26, 27].

이종 표현은 단어, 문장, 문서, 개념 등에서 두 개 이상의 타입들을 노드로 표현하는 방식이다. 이 방식의 가장 흔한 형태가 이분 그래프(bipartite graph)이다[8, 11, 21, 28]. 예를 들어 그림 2는 문서와 개념을 이분 그래프로 표현한 예를 보여준다. 이종 표현의 또 다른 예는 잠재적 의미 분석(LSA: Latent Semantic Analysis)에서 찾아볼 수 있다[18]. LSA는 문서와 그 문서에 출현하는 단어들을 행렬로 표현한 후 특이값 분해(SVD: Singular Vector Decomposition)을 통해 차원을 축소하여 단어 간 혹은 문서 간 잠재적 관련성을 탐색하는 방식이다. LSA에서는 문서와 단어 간 관계를 행렬로 표현하는데 결국 이는 문서와 단어 간의 이분 그래프와 동일한 형태이므로 이종 표현 방식 중 하나로 분류될

수 있다.

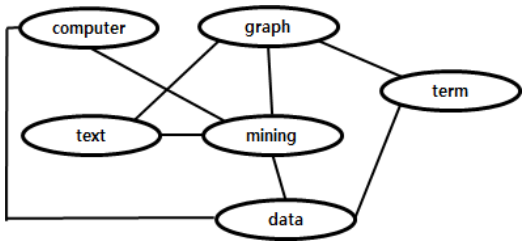

그림 1. 동종표현 모델의 예
Fig. 1. An example of homogeneous representation model

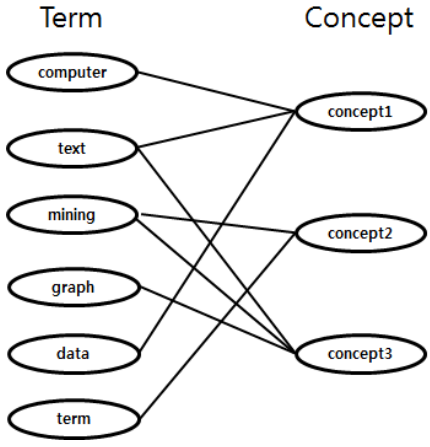

그림 2. 이분 그래프의 예
Fig. 2. An example of bipartite graph

(2) Weighted vs. Unweighted

weighted는 노드에 가중치가 부여된 형태를 말하며, unweighted는 그렇지 않은 그래프 형태를 말한다. 일부의 연구를 제외하고는 대부분 weighted를 가정하는데, 가중치는 그래프 내에서 해당 노드의 중요도를 나타낸다. 예를 들어 단어들의 공기 그래프에서 노드의 가중치는 간단히 해당 단어의 TF-IDF 값을 계산하여 나타낼 수 있다. 하지만 이 경우는 벡터공간 모델에서의 가중치와 동일하므로 그래프를 이용한 장점을 살릴 수 없게 된다. 따라서 대부분의 연구에서는 노드에 연결된 간선의 수나 간선들의 가중치 혹은 해당 노드에 연결된 이웃 노드들의 가중치를 이용하여 간접적으로 노드의 중요도를 계산하는 방식을 사용한다. 이러한 방식의 대표적인 예로 PageRank[29]를 들 수 있다.

나. 간선의 표현방식

간선은 노드 간에 관련성을 가질 때 이들을 연결하는데 이용된다. 간선은 그 형태에 따라 세 가지 종류의 분류체계를 갖는데 우선 방향성을 갖느냐의 여부에 따라 directed 또는 undirected로 나눌 수 있고, 간선에 가중치가 부여되느냐에 따라 weighted 또는 unweighted로 나눌 수 있다. 마지막으로 간선에 레이블이 부여되느냐에 따라 labeled 또는 unlabeled로 분류될 수 있다.

(1) Directed vs. Undirected

directed는 노드간의 순서나 상호간의 역할이 중요한 특징이 될 때 사용된다. 예를 들어 문서나 문장 내에 단어가 출현한 순서를 표현하고자 할 때는 directed 표현방식을 사용한다[3, 9, 16, 19, 23]. 문장내의 단어들에 대해서 주어-동사-목적어와 같이 문법적 의존성을 표현할 때도 사용되며[6, 13], 트리 형태의 그래프 표현방식도 directed로 분류될 수 있다[24].

반면에 undirected는 관련성이 있으나 그 순서나 상호간의 역할이 부여되지 않을 때 사용된다. 가장 흔한 경우가 단어 간의 공기 정보를 표현할 때 undirected 방식을 사용한다[4, 20, 22].

(2) Weighted vs. Unweighted

weighted는 노드간의 연관성 정도를 점수로 표현하고자 할 때 사용된다. 예를 들어, 단어 간의 공기 그래프에서 간선에 연결된 두 단어가 동시에 출현하는 빈도수에 따라 가중치를 부여할 수 있다[3, 9]. 또 다른 예로는 간선에 연결된 두 단어의 거리로서 가중치를 부여할 수 있다[20]. 즉, 두 단어의 거리가 멀수록 연관성이 떨어지므로 가중치를 작게 부여하고 그 반대의 경우 가중치를 높게 부여할 수 있다. 또 다른 예로는 문장을 노드로 표현하는 그래프에서 문장 간의 유사도에 따라 가중치를 부여하는 경우도 있다[14, 15, 25].

반대로 노드간의 연관성은 있으나 그 정도를 정량적으로 표현할 필요가 없는 경우에는 unweighted 방식을 사용한다[6, 12, 13, 19].

(3) Labeled vs. Unlabeled

일부 그래프 표현 모델에서는 간선에 레이블의 표현하는 경우도 있다[5, 6, 7, 13, 19, 24]. 레이블은 간선의 역할을 부여할 때 사용되는데, 많은 경우 단어와 단어 사이

의 관계를 표현한다. 예를 들어 [13]에서는 명사로 구성된 노드 집합에서 주어(subject)와 목적어(object)의 관계를 표현할 때 간선에 동사(verb)를 레이블로 부여한다. 또한 [6]에서는 문장을 파싱(parsing)한 형태의 트리로 구성하는데, 각 단어의 품사(PoS: Part of Speech)를 간선의 레이블로 표현한다. 이와 같이 레이블을 부여하는 경우는 대부분 문서의 문법적 요소나 구조적 형태(예를 들어 html 또는 xml 문서)를 그래프로 표현할 때 많이 사용한다. 그 이외의 경우는 대부분 unlabeled 방식을 사용한다.

2. 그래프의 내용(contents)에 따른 분류

그래프 기반 텍스트 모델을 분류하는 또 다른 접근 방법은 그래프가 표현하고자하는 내용에 따라 분류하는 것이다. 그래프의 내용은 크게 세 가지로 분류할 수 있는데 첫째는 노드로 표현된 요소들에 대한 공기 또는 유사성을 표현하기 위한 모델이 있고, 또 하나는 노드 간의 문법적 연관성을 표현하는 모델이 있다. 마지막으로는 노드의 의미적 연관성을 표현하기 위한 모델이 있다.

가. 공기 또는 유사성 표현 모델

이 모델은 기존 연구에서 가장 많이 사용되고 있는 방식으로 단어 간의 공기 정보나 문장 간의 유사도 등을 표현한다[3, 4, 9, 12, 17, 20, 22, 23, 26, 27]. 이 모델은 다른 모델에 비해 상대적으로 단순하며, 구축비용도 적게 든다. 또한 기존의 그래프 마이닝(graph mining) 분야에서 제안된 다양한 알고리즘에 대한 적용이 쉽다. 마지막으로 이 모델은 언어에 독립적(language independent)이다. 즉, 영어를 대상으로 제안된 알고리즘들은 한국어를 비롯한 기타 언어에도 동일하게 적용될 수 있다. 그러나 단순성으로 인해 표현력이 다른 모델에 비해 약하다는 단점을 안고 있다.

나. 문법적 연관성 표현 모델

문법적 연관성을 표현하는 모델에서는 자연어 처리 기법을 이용하여 노드를 품사의 타입별로 구분하고 이를 간선의 레이블로 표현함으로써 노드간의 의존성(dependency)을 나타낼 수 있다[6, 13]. 이 기법은 문장의 구조를 자세히 표현할 수 있다는 장점이 있으나 그래프의 복잡도가 증가하여 계산 비용이 많이 든다는 단점이 있다. 또한 문법적 규칙을 잘 지키지 않는 SNS 문서와 같은 환경에서는 오류의 가능성이 매우 커지게 된다. 특히 자유도가 높은 한글 문서의 경우에는 그 가능성이 더욱 높다.

다. 의미적 연관성 표현 모델

의미적 연관성에 대한 표현하는 방법은 개념을 노드로 표현하는 것이다. 대표적인 예가 문서와 개념 간의 관계를 이분 그래프로 표현한 모델이다[11, 28]. 여기서는 문서에 나타난 중요 단어들을 개념으로 취급하여 문서와 개념 간의 연관 관계를 이분 그래프로 표현한다. 또 다른 예는 개념 트리를 구성하는 것으로 문서나 단어를 포함하는 대표적 개념을 선정하고 개념과 개념간의 관계를 트리 형태로 표현한다[10, 16]. 이와 같이 개념을 노드로 표현하는 방식에서는 사전에 이미 구축된 개념 집합이 존재해야하는데 대표적으로 [16, 28]에서는 Wikipedia를 이용하였고, [10]에서는 WordNet을 이용하였다. 반면에 [8]에서도 문서와 개념간의 이분 그래프를 이용하지만 별도의 개념집합을 사용하지 않고, 잠재적 의미 분석을 통해 선정된 문서내의 주요 단어를 개념으로 취급하였다.

Ⅳ. 그래프 기반 텍스트 마이닝에서의 대표적 기술

본 장에서는 그래프 기반 텍스트 표현 모델 연구들에서 사용되었던 주요 기술이나 알고리즘들을 소개한다. 일부 연구에서는 자체적으로 개발한 알고리즘을 사용하는 경우도 있으나, 많은 경우에는 기존에 제안되었던 잘 알려진 기술들을 직접적으로 또는 상황에 맞게 응용해서 사용한다. 응용분야에 관계없이 그래프 기반 표현 모델에서는 공통적으로 노드나 간선의 가중치를 계산하는 방법과 그래프 구조 핵심이 되는 서브 그래프를 탐색하는 기술이 요구된다.

1. 노드나 간선의 가중치 계산 기술

노드의 가중치를 계산하는 가장 간단한 방법은 노드가 단어일 경우 해당 단어의 출현 빈도나 TF-IDF를 계산하는 것이다. 간선의 경우에는 연결된 노드들에 대한 공기 빈도로 표현할 수 있으며, 노드가 문장일 경우

에는 공기 정보를 표현할 수가 없으므로, 문장과 문장의 유사도를 계산하는 방식을 사용하기도 한다. 예를 들어 문장과 문장의 유사도는 코사인 유사도(cosine similarity)나 유클리디언 거리를 이용하여 계산할 수 있다. 그 이외에 제안된 주요 기술들은 다음과 같다.

가. PageRank

PageRank는 잘 알려진 그래프 랭킹 알고리즘 중의 하나이다. PageRank가 개발된 초창기에는 Google과 같은 World Wide Web 환경에서 검색을 위한 웹 페이지들의 랭킹을 위해 개발되었으며[29], 이후에는 그래프 구조에서 노드에 대한 랭킹에 광범위하게 응용되고 있다. PageRank에서 노드 V_i에 대한 가중치는 이와 간선으로 연결된 다른 노드들의 개수와 중요도로 평가하는데, V_i의 가중치 $PR(V_i)$는 다음과 같이 정의된다.

$$PR(V_i) = (1-d) + d^* \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \quad (5)$$

이 식에서 $In(V_i)$는 V_i를 가리키는 노드들의 집합을 의미하며, $Out(V_i)$는 V_i가 가리키는 노드들의 집합을 타나낸다. 그리고 d는 감쇠율(damping factor)로 0부터 1까지의 값을 갖는데 일반적으로 0.85로 설정된다[29]. PageRank는 [4, 15, 26, 27]등에서 문서 분류나 요약 등의 목적으로 노드들의 가중치 계산을 위해 사용하였다.

나. TextRank와 LexRank

TextRank 알고리즘은 PageRank에 바탕을 둔 그래프 기반 랭킹 알고리즘이다[17]. 다만 차이점은 TextRank는 기본적으로 undirected 그래프를 가정하며, 유사도를 계산하기 위해 간선에 가중치가 부여된 그래프까지 고려하였다. 예를 들어 간선에 가중치가 부여된 그래프의 경우 노드의 가중치는 다음과 같은 수식으로 계산된다.

$$PR(V_i) = (1-d) + d^* \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\left| \sum_{V_k \in Out(V_j)} w_{jk} \right|} PR(V_j) \quad (6)$$

이 식에서 w_{ij}는 V_i와 V_j를 잇는 간선의 가중치를 의미한다. TextRank 알고리즘은 기본적으로 문서 요약을 위해 개발되었으나[17], 노벨티 탐색에도 응용되었다[20].

TextRank와 유사한 방법으로 LexRank 알고리즘도 제안되었다[14]. 두 알고리즘의 차이점은 LexRank에서는 대용량 문서 요약을 위한 일부 방법 중 하나로 이 기법을 이용하였다는 것이며, TextRank와는 달리 다중 문서(multi-document) 요약에 적용되었다는 점이다. 그 이외에 두 알고리즘은 거의 동일다고 볼 수 있다.

다. HITS

HITS(Hyperlink-Induced Topic Search) 알고리즘은 웹 문서들의 랭킹을 위해 개발된 방법으로, PagerRank에 앞서 제안되었으며 PageRank의 개발에 큰 영향을 주었다[30]. 이 기법의 기본 아이디어는 특정 페이지가 중요한 페이지들에 의해 링크가 된다면, 해당 페이지의 랭크를 높게 부여하는 것이다.

이 알고리즘은 텍스트 그래프 모델에서 노드의 가중치를 계산하는 데 응용될 수 있다[12, 21, 26, 27]. 구체적으로 HITS에서는 각 노드에 대해서 authorities 점수와 hubs 점수를 부여하는데, authorities는 들어오는 링크가 많은 노드에 대한 점수이며, hubs 노드는 밖으로 나가는 링크가 많은 노드에 대한 점수이다. 이들 각각은 다음과 같이 계산된다.

$$HITS_A(V_i) = \sum_{V_j \in In(V_i)} HITS_H(V_j)$$
$$HITS_H(V_i) = \sum_{V_j \in Out(V_i)} HITS_A(V_j) \quad (7)$$

최종적으로 노드에 대한 가중치는 authorities와 hubs 점수의 합이나 평균 등의 계산을 통해 얻을 수 있다.

라. PMI

PMI(Pointwise Mutual Information)은 두 개체간의 연관 정도를 측정하는 방법 중의 하나로 그래프에서 간선의 가중치를 계산하기 위해서 사용된다[11, 20]. 예를 들어 단어에 대한 공기 그래프에서 $P(i)$와 $P(j)$를 각각 단어 w_i와 w_j가 문서에 포함될 확률이고, $P(i, j)$를 두 단어가 동시에 문서에 포함될 확률이라면 w_i와 w_j에 대한 PMI는 다음과 같이 계산된다.

$$PMI_{ij} = \log_2 \frac{P(i,j)}{P(i)P(j)} \quad (8)$$

따라서 두 단어가 독립적이면 0의 값을 갖게 되며, 0보다 클수록 연관성이 높아진다고 볼 수 있다.

마. 빈발 항목 마이닝

빈발 항목 마이닝(Frequent Itemset Mining)은 association rule mining이라고도 하며, 동시에 출현하는 여러 항목 집합에서 동시에 출현하는 항목 집합을 탐색하는 기술이다. 이 기법은 두 노드가 빈발 항목에 포함되면, 해당 값으로 간선의 가중치를 부여하는 방법으로 사용된다[4]. 또한 이 기술은 빈발 서브그래프(frequent subgraph) 탐색하는 기술로도 활용된다[10].

2. 서브 그래프 탐색 기술

그래프 기반 텍스트 표현 모델은 기본적으로 그래프 마이닝에서 제안된 다양한 분석 기술들을 이용할 수 있다는 장점이 있다. 본 논문에서는 지금까지의 연구에서 적용된 그래프 마이닝 분석 기술들을 정리한다.

가. 빈발 서브그래프 마이닝

빈발 서브그래프 마이닝(FSM: Frequent Subgraph Mining)은 하나 단일 혹은 다중 그래프 구조에서 빈발하게 나타나는 서브 그래프를 탐색하는 기술이다. 이 기법은 그래프 마이닝의 핵심 기술로 화학, 생물학, 웹 환경 등에 광범위하게 응용되고 있다[31].

그래프 기반 텍스트 표현 모델에서는 주로 문서 분류나 군집화 목적으로 사용된다[6, 10]. 문서 분류에서는 학습 데이터로 주어진 텍스트 그래프에서 빈발 서브그래프를 탐색하여 이들을 클래스에 대한 특징으로 정하는데 이용될 수 있다[6]. 또한 빈발 서브그래프를 탐색한 후 이들의 유사도를 이용하여 문서들을 군집화하는데 이용하기도 한다[10].

나. 최대 공통 그래프 탐색 기법

최대 공통 그래프(MCS: Maximum Common Subgraph)는 그래프 이론(graph theory)에서 나온 개념으로, 다중 그래프에서 공통적으로 갖는 최대의 서브그래프를 탐색하는 기술이다. 그래프 기반 텍스트 모델에서 MCS는 주로 문서 분류를 위해 사용된다[3, 7]. MCS의 가장 큰 문제점은 계산 비용이 NP-complete라는데 있다. 그러나 대부분의 연구에서는 MCS에 대한 구체적인 알고리즘을 제시하지 않던지[23], 아니면 단순화된 그래프 모델을 사용하여 다항시간(polynomial time) 내에 문제를 해결하는 방안을 제시하고 있다[7].

다. SimRank

SimRank는 그래프 구조에서 두 노드의 유사도를 구하기 위한 방법으로 제안되었다[28, 32]. SimRank는 두 노드가 유사한 노드들에 의해 공통적으로 연관되어 있다면 그들은 서로 유사하다는 아이디어에 착안하여 개발되었다. 두 노드 a, b에 대한 유사도 $s(a, b)$는 다음과 같이 계산된다. 우선 a와 b가 동일하면 $s(a, b)$=1이 된다. 그 이외에는 다음과 같이 계산된다.

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) \quad (9)$$

이 식에서 C는 0과 1 사이의 상수이며, $I(v)$는 노드 v로 들어오는 링크로 연결된 노드들의 집합이다. 또한 $I_i(v)$는 $1 \le i \le |I(v)|$ 인 개별 노드를 의미한다.

V. 국내의 그래프기반 텍스트 마이닝 동향

국내에서도 그래프 기반 텍스트 마이닝에 대한 다양한 연구가 진행되었으며, 응용분야도 외국의 사례와 유사하게 분류, 요약, 검색, 군집화 등을 목적으로 연구되었다[33, 34, 35, 36, 37, 39, 40, 41]. 그러나 국내의 연구사례를 살펴보면 방법론적 측면에서 외국의 사례와 큰 차이를 보이지 않고 있다. 즉, 새로운 알고리즘에 대한 제시보다는 대부분 외국에서 제안된 기초 알고리즘에 의존하고 있다. 더구나 한글 문서를 대상으로 한 연구도 매우 드물다. 이는 한글을 대상으로 한 테스트 문서의 부재가 가장 큰 원인으로 파악되고 있다. 따라서 한글을 대상으로 한 그래프 기반 텍스트 마이닝의 발전을 위해서는 한글이 갖고 있는 고유의 특징을 반영한 모델을 개발하고, 성능을 검증 할 수 있는 공인된 테스트 문서 집합의 개발이 우선되어야한다.

표 1. 대표적인 그래프기반 텍스트 마이닝 연구
Table1. Representative researches of graph-based text mining

참고 문헌	응용 분야	그래프 구조		그래프 내용	적용된 주요 알고리즘
		노드	간선		
3	classification	homo(term)	directed, weighted, unlabeled	co-occurrence	MCS
4	classification	homo(term)	undirected, weighted, unlabeled	co-occurrence	PageRank
5	classification	homo(term)	directed, unweighted, labeled/unlabeled	co-occurrence, syntax	
6	classification	hetero (term+PoS)	directed, unweighted, labeled	syntax, semantic	FSM
7	classification	homo(term)	directed, unweighted, labeled	co-occurrence	MCS
8	classification	hetero (doc+concept)	undirected, weighted, unlabeled	bipartite graph	
9	clustering	homo(term)	directed, weighted, unlabeled	co-occurrence	
10	clustering	hetero (term+concept)	directed, unweighted, unlabeled	semantic tree	FSM
11	clustering	hetero (doc+concept)	undirected, unweighted, unlabeled	bipartite graph	PMI
12	summarization	homo(term)	directed, unweighted, unlabeled	co-occurrence	HITS
13	summarization	homo(term)	directed, unweighted, labeled	syntax	
14	summarization	homo(sentence)	undirected, weighted, unlabeled	similarity	LexRank
15	summarization	homo(sentence)	undirected, weighted, unlabeled	similarity	PageRank
16	keyword extraction,	hetero (term+concept)	directed, weighted, unlabeled	semantic tree	
17	keyword extraction, summarization	homo(term or sentence)	directed/undirected, weighted/unweighted, unlabeled	co-occurrence, similarity	TextRank
19	search	homo(term)	directed, unweighted, labeled	co-occurrence, syntax	
20	novelty detection	homo(term)	undirected, weighted, unlabeled	co-occurrence	PMI, TextRank
21	opinion mining	hetero (term pair+doc)	undirected, unweighted, unlabeled	bipartite graph	HITS
22	search	homo(term)	undirected, weighted, unlabeled	co-occurrence	
23		homo(term)	directed, weighted, labeled	co-occurrence	MCS
24	opinion mining	homo(term)	directed, weighted, labeled	semantic tree	
25	summarization	homo(sentence)	undirected, weighted, unlabeled	similarity	
26	summarization	homo(sentence)	directed/undirected, weighted/unweighted, unlabeled	similarity	PageRank, HITS
27	summarization	homo(sentence)	directed/undirected, weighted, unlabeled	similarity	PageRank, HITS
28	classification clustering	hetero (doc+concept)	directed/undirected, weighted, unlabeled	bipartite graph	SimRank

* 이 표에서 약어로 표현된 단어들의 의미는 다음과 같다.
homo(homogeneous), hetero(heterogeneous), doc(document), Pos(Part of Speech)

VI. 결론 및 향후 발전 방향

본 논문에서는 기존에 제안되었던 그래프 기반 텍스트 표현 모델의 방법과 종류들을 제시하였다. 우선 그래프 표현 모델에 있어서 노드와 간선의 종류를 그 특성에 따라 나누었다. 노드에 대해서는 노드를 표현하는 객체의 다양성에 따라 동종 표현과 이종 표현으로 구분하였고, 노드에 가중치를 부여 여부에 따라 weighted와 unweighted로 나누었다. 그러나 대부분의 연구에서는 weighted를 가정하므로 이러한 구분은 큰 의미는 없다고 하겠다. 간선에 대해서는 directed 또는 undirected, weighted 또는 unweighted, labeled 또는 unlabeled로 구분하였다. 또한 그래프의 내용에 따라 각각 공기 또는 유사성 표현모델, 문법적 연관성 표현 모델, 마지막

으로 의미적 연관성 표현 모델로 구분하였다.

이외에도 본 논문에서는 그래프 기반 텍스트 표현 모델에서 서브 그래프를 탐색하기 위한 여러 가지 기법들을 소개하였다. 이러한 기법들은 대부분 그래프 마이닝 연구에서 제안되었던 것들이다. 대표적으로 빈발 서브그래프 마이닝, 최대 공통 그래프 탐색, SimRank 등을 정리하였다. 이외에도 그래프 마이닝에서 제안된 다양한 기술들을 그래프 기반 텍스트 표현 모델에 적용할 수 있다.

표 1은 본 논문에서 제시한 그래프 기반 텍스트 모델에 관련된 대표적인 연구들을 정리한 것이다. 이 표에서 첫 번째 컬럼은 참고문헌 번호를 나타내며, 두 번째 컬럼은 해당 연구의 최종 적용 분야를 나타낸다. 세 번째와 네 번째는 각각 노드와 간선의 특징을 나타내고, 다섯 번째 컬럼은 그래프가 표현하고자하는 내용을 나타낸다. 마지막 컬럼은 해당 연구에서 적용한 대표적인 알고리즘을 나타내는데, 이 표에서는 잘 알려진 알고리즘만을 기술하였다.

지금까지 살펴본 바와 같이 그래프 기반 텍스트 표현 모델은 텍스트 분석을 위한 목적과 응용분야(분류, 군집화, 요약, 검색 등)에 따라 다양하고 독립적인 그래프 모델을 사용하고 있다. 이는 역으로 다양한 목적에 적용 가능한 표준화된 그래프 모델을 제시한 사례가 부재하다는 의미이기도 하다. 예를 들어 문서 분류를 위해 제안된 그래프 모델은 군집화나, 요약, 검색을 위한 방법에 응용되기 어려운 점이 있다. 따라서 향후 연구에서는 문서 표현을 위한 체계화된 그래프 모델의 개발이 요구된다. 이러한 개발이 이루어진다면 이를 기반으로 하여 문서분류, 군집화, 요약, 검색 등 기존의 다양한 문서 분석기술에 응용할 수 있을 것으로 기대된다.

References

[1] G. Salton, A. Wong, and C. S. Yang , "A Vector Space Model for Automatic Indexing," Communications of the ACM, Vol. 18, Vo. 11, pp. 613 - 620, 1975.

[2] G. Salton and M. J. Mcgill, Introduction to Moderm Information Retrieval, McGraw-Hill, New York, 1983.

[3] J. Wu, Z. Xuan, and D. Pan, "Enhancing Text Representation for Classification Tasks with Semantic Graph Structures", International Journal if Innovative Computing, Information Control, Vol. 7, No. 5(B), pp. 2689-2698, 2011.

[4] W. Wang, D. B. Do, and X. Lin, "Term Graph Model for Text Classification", Proceedings of the First international conference on Advanced Data Mining and Applications, pp. 19-30, 2005.

[5] K. Valle and P. Ozturk, "Graph-Based Representation for Text Classification", India-Norway Workshop on Web Concepts and Technologies, 2011.

[6] C. Jiang F. Coenen, R. Sanderson, and M. Zito, "Text Classification Using Graph Mining-Based Feature Extraction", Knowledge-Based Systems, Vol. 23, No. 4, pp. 302-308, 2009.

[7] A. Schenker, M. Last, H. Bunke, and A. Kandel, "Classification of Web Documents Using a Graph Model", 2003. Proceedings. Seventh International Conference on Document Analysis and Recognition, pp. 240-244, 2003.

[8] R. Chau, A. C. Tsoi, M. Hagenbuchner, and V. C.S. Lee, "A Concept Graph for Text Structure Mining", Proceedings of the Thirty-Second Australasian Conference on Computer Science, Vol 91, pp. 141-150, 2009.

[9] K. M. Hammouda and M S. Kamel, "Document Similarity Using a Phrase Indexing Graph Model", Knowledge and Information Systems, Vol. 6, No. 6, pp. 710-727, 2006.

[10] M. S. Hossain, R. A. Angryk, "GDClust: A Graph-Based Document Clustering Technique", Proceedings of Seventh IEEE International Conference on Data Mining Workshops, pp. 417-422, 2007.

[11] I. Yoo, X. Hu, and I.-Y. Song, "Integration of Semantic-based Bipartite Graph Representation and Mutual Refinement Strategy for Biomedical Literature Clustering", Proceedings of the 12th ACM SIGKDD international conference on

Knowledge discovery and data mining, pp. 791-796, 2006.

[12] M. Litvak and M. Last, "Graph-Based Keyword Extraction for Single-Document Summarization", Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, pp. 17-24, 2008.

[13] J. Leskovec, M. Grobelnik, and N. Milic-Fraying, "Learning Semantic Graph Mapping for Document Summarization", Proceedings of the ECML/PKDD-2004 Workshop on Knowledge Discovery and Ontologies. 2005.

[14] G. Erkan and D. R. Radev, "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Vol. 22, No. 1, pp. 457-479, 2004.

[15] S. Hariharan and R. Srinivasan, "Studies on Graph based Approaches for Single and Multi Document Summarizations", International Journal of Computer Theory and Engineering, Vol. 1, No. 5, pp. 1793-8201, 2009.

[16] C. A. Chahine, N. Chaignaud, JHP Kotowicz, and JP Pecuchet, "Context and Keyword Extraction in Plain Text Using a Graph Representation", Proceedings of the 2008 IEEE International Conference on Signal Image Technology and Internet Based Systems, pp. 692-696, 2008.

[17] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts", Proceedings of International Conference on Empirical Methods in Natural Language Processing, 2004.

[18] S. T. Dumais, "Latent Semantic Analysis", Annual Review of Information Science and Technology, Vol. 38, No. 1, pp. 188-230, 2004

[19] S. Hensman, "Construction of Conceptual Graph Representation of Texts", Proceedings of the Student Research Workshop at HLT-NAACL, pp. 49-54, 2004.

[20] M. Gamon, "Graph-Based Text Representation for Novelty Detection", Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing, pp. 17-24, 2006.

[21] B. Li, L. Zhou, S. Feng, and K.-F. Wong "A Unified Graph Model for Sentence-Based Opinion Retrieval" Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1367-1375, 2010.

[22] J. Tomita, H. Nakawatase, and M. Ishii, "Graph-Based Text Database for Knowledge Discovery", Proceedings of the 13th international World Wide Web conference, pp. 454-455, 2004.

[23] F. Zhou, F. Zhang, and B. Yang, "Graph-Based Text Representation Model and its Realization", Proceedings of International Conference on Natural Lan guage Processing and Knowledge Engineering, pp. 1-8, 2010.

[24] Y. Wu, Q. Zhang X. Huang, and L Wu, "Structural Opinion Mining for Graph-based Sentiment Representation", Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1332-1341, 2011.

[25] X. Wan and J. Yang, "Improved Affinity Grapg Based Multi-Document Summarization", Proceedings of the Human Language Technology Conference of the NAACL, pp. 181-184, 2006.

[26] R. Mihalcea, "Graph-Based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization", Proceedings of 3rd International Conference on Emerging Trends in Engineering and Technology(ICETET), pp. 516-519, 2010.

[27] R. Mihalcea and P. Tarau, "A Language Independent Algorithm for Single and Multiple Document Summarization", Proceedings of International Joint Conference on Natural Language Processing, 2005.

[28] L. Zhang, C. Li, J. Liu, and H. Wang, "Graph-Based Text Similarity Measurement by Exploiting Wikipedia as Background Knowledge", World Academy of Science, Engineering and Technology, Issue 59, pp. 1548-1553, 2011.

[29] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine", Proceedings of the seventh International Conference on World Wide Web 7, pp. 107-117, 1998.
[30] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Journal of ACM, Vol. 45, No. 5, pp. 605-632, 1999.
[31] C. Jiang, F. Coenen, and M. Zito, "A Survey of Frequent Subgraph Mining Algorithm", The Knowledge Engineering Review, Vol. 28, Issue 1, pp. 75-105, 2012.
[32] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity", Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 538-543, 2002
[33] W.-S. Bae and J.-W Cha, "Text Categorization Using TextRank Algorithm", Journal of KIISE, Vol. 16, No. 1, pp. 110-114, 2010.
[34] J. H. Lyu and S. C. Park, "Document Summarization Method Using Complete Graph", Journal of Korea Society of Industrial Information Systems, Vol. 10, No. 2, pp. 26-31, 2005.
[35] H. K. Bae, H. Park, S. Lee, and K. Kim, "Improved Concept-based Search System Using HITS Algorithm on Conceptual Graph", Proceedings of KIISE conference, pp. 470-472, 2003.
[36] S. Cho and K. Lee, "Query Expansion Based on Word Graphs Using Pseudo Non-Relevant Documents and Term Proximity", Journal of KIPS, Vol 19B, No. 3, pp. 189-194, 2012.
[37] W. M. Song, Y. Kim, E.-J. Kim, and M. Kim, "A Document Summarization System Using Dynamic Connection Graph", Journal of KIISE, Vol. 36, No. 1, pp. 62-69, 2009.
[38] http://en.wikipedia.org/wiki/Vector_space_mode
[39] M. Hwang, D. Choi, and P. Kim "A Context Information Extraction Method according to Subject for Semantic Text Processing", Journal of Korean Institute of Information Technology, vol. 8, No. 11, pp. 197-204, 2010.
[40] J. Shim, H. C. Lee, "The Development of Automatic Ontology Generation System Using Extended Search Keywords" Journal of the Korea Academia-Industrial cooperation Society, Vol. 11, no. 6, 2009.
[41] J. Chang, "Efficient Retrieval of Short Opinion Documents Using Learning to Rank", Journal of the Institute of Internet, Broadcasting and Communication, Vol. 13, No. 4, Aug., 2013.

※ 본 연구는 한성대학교 교내학술연구비 지원과제임.

저자 소개

장 재 영(정회원)

- 1992년 : 서울대학교 계산통계학과 (이학사)
- 1994년 : 서울대학교 계산통계학과 (이학석사)
- 1999년 : 서울대학교 계산통계학과 (이학박사)
- 2000년~현재 : 한성대학교 컴퓨터공학과 교수

<주관심분야 : 데이터베이스, 정보검색, 데이터마이닝>