그래프 기반 텍스트 마이닝의 연구 동향†

장재영*, 한종빈**, 좌태빈***
한성대학교 컴퓨터공학과
e-mail: jychang@hansung.ac.kr*
gkswhdqls001@naver.com**
jtb2005@naver.com***

Research Trends of Graph-Based Text Mining

Jae-Young Chang*, Jong Bin Han**, Tae Bin Jwa*
Dept. of Computer Science, Hansung University

요 약

텍스트 마이닝은 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 주제별로 제시한다.

1. 서론

텍스트 마이닝(text mining)은 비정형 문서를 대상으로 한 데이터 마이닝(data mining)의 한 분야로서 문서에 숨겨진 고급 지식들을 탐색하는 분야이다. 텍스트 마이닝에서 텍스트에 대한 표현 모델로서 지금까지 가장 많이 사용되고 있는 것은 벡터공간 모델(VSM: Vector Space Model)이다[1]. 벡터공간 모델에서는 문서에 출현하는 주요 단어와 그들의 가중치(weight)를 벡터 형태로 표현한다. 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, word), 문장(sentence), 구(phrase), 개념(concept) 등의 공기(co-occurrence) 또는 기타 관계(relation) 정보를 활용하여 문서의 특징을 보다 정밀하게 표현할 수 있는 장점이 있다.

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 체계적으로 정리한다.

2. 벡터공간 모델

벡터공간 모델에서는 단어를 하나의 차원으로 표현하고, 이를 이용하여 문서를 n차원 공간의 하나의 점(point)로 표현한다. 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[2].

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

3. 그래프 기반 텍스트 모델의 종류

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.

3.1 그래프 구조에 따른 분류

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

$$G = \{V, E\}$$

† 이 논문은 2011년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임.(과제번호: NRF-2011-0022445).

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.

3.1.1 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 하나의 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.

동종 표현 vs. 이종 표현

동종 표현에서 가장 흔하게 나타나는 방식이 단어를 노드로 표현하는 것이다[3-8]. 여기서는 많은 경우 단어 간의 공기 정보로 그래프로 표현한다. 공기 정보란 단어들이 동시에 출현하는 것을 나타내는 것으로 하나의 문서나 문장 내에 두 단어가 동시에 나타나면 이를 간선을 연결하는 형태를 말한다. 이 밖에도 단어 간의 문법적 연관성이나 의미적(semantic) 유사성에 따라 그래프로 표현하기도 한다[10]. 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[12, 17, 18].

이종 표현은 단어, 문장, 문서, 개념 등에서 두 개 이상의 타입들을 노드로 표현하는 방식이다. 이 방식의 가장 흔한 형태가 이분 그래프(bipartite graph)이다[11, 15].

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. 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[19]를 들 수 있다.

3.1.2 간선의 표현방식

간선은 노드 간에 관련성을 가질 때 이들을 연결하는데 이용된다. 간선은 그 형태에 따라 세 가지 종류의 분류 체계를 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.

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[4]. 문장내의 단어들에 대해서 주어-동사-목적어와 같이 문법적 의존성을 표현할 때도 사용되며[9, 10], 트리 형태의 그래프 표현방식도 directed로 분류될 수 있다[13].

반면에 undirected는 관련성이 있으나 그 순서나 상호간의 역할이 부여되지 않을 때 사용된다. 가장 흔한 경우가 단어 간의 공기 정보를 표현할 때 undirected 방식을 사용한다[6, 18].

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, 4]. 또 다른 예로는 간선에 연결된 두 단어의 거리로서 가중치를 부여할 수 있다. 반대로 노드간의 연관성은 있으나 그 정도를 정량적으로 표현할 필요가 없는 경우에는 unweighted 방식을 사용한다[5, 8, 9, 10].

Labeled vs. Unlabeled

일부 그래프 표현 모델에서는 간선에 레이블의 표현하는 경우도 있다[5, 7, 9, 10, 13]. 레이블은 간선의 역할을 부여할 때 사용되는데, 많은 경우 단어와 단어 사이의 관계를 표현한다. 예를 들어 [10]에서는 명사로 구성된 노드 집합에서 주어(subject)와 목적어(object)의 관계를 표현할 때 간선에 동사(verb)를 레이블로 부여한다. 또한 [9]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.

3.2 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 의

미적 연관성을 표현하기 위한 모델이 있다.

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, 6, 8, 16, 17, 18]. 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.

문법적 연관성 표현 모델

문법적 연관성을 표현하는 모델에서는 자연어 처리기법을 이용하여 노드를 품사의 타입별로 구분하고 이를 간선의 레이블로 표현함으로써 노드간의 의존성(dependency)을 나타낼 수 있다[9, 10]. 이 기법은 문장의 구조를 자세히 표현할 수 있다는 장점이 있으나 그래프의 복잡도가 증가하여 계산 비용이 많이 든다는 단점이 있다.

의미적 연관성 표현 모델

의미적 연관성에 대한 표현하는 방법은 개념을 노드로 표현하는 것이다. 대표적인 예가 문서와 개념 간의 관계를 이분 그래프로 표현한 모델이다[15]. 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. 이와 같이 개념을 노드로 표현하는 방식에서는 사전에 이미 구축된 개념 집합이 존재해야하는데 대표적으로 [11]에서는 Wikipedia를 이용하였다.

4. 결론 및 향후 발전방향

논문에서는 기존에 제안되었던 그래프 기반 텍스트 표현 모델의 방법과 종류들을 제시하였다. 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.

지금까지 살펴본 바와 같이 그래프 기반 텍스트 표현 모델은 텍스트 분석을 위한 목적과 응용분야에 따라 다양하고 독립적인 그래프 모델을 사용하고 있다. 이는 역으로 다양한 목적에 적용 가능한 표준화된 그래프 모델을 제시한 사례가 부재하다는 의미이기도 하다. 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.

참고문헌

[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] http://en.wikipedia.org/wiki/Vector_space_model

[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] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Wu, Q. Zhang X. Huang, and L Wu, "Structural Opinion Mining for Graph-based Sentiment

<표 1> 그래프 기반 텍스트 마이닝의 기존 연구 특징*

참고 문헌	응용 분야	그래프 구조		그래프 내용
		노드	간선	
3	classification	homo(term)	directed, weighted, unlabeled	co-occurrence
4	clustering	homo(term)	directed, weighted, unlabeled	co-occurrence
5	search	homo(term)	directed, unweighted, labeled	co-occurrence, syntax
6	classification	homo(term)	undirected, weighted, unlabeled	co-occurrence
7	classification	homo(term)	directed, unweighted, labeled/unlabeled	co-occurrence, syntax
8	summarization	homo(term)	directed, unweighted, unlabeled	co-occurrence
8	classification	hetero(term+PoS)	directed, unweighted, labeled	syntax, semantic
10	summarization	homo(term)	directed, unweighted, labeled	syntax
11	classification clustering	hetero(doc+concept)	directed/undirected, weighted, unlabeled	semantic (bipartite graph)
12	summarization	homo(sentence)	undirected, weighted, unlabeled	similarity
13	opinion mining	homo(term)	directed, weighted, labeled	semantic tree
14	summarization	homo(sentence)	undirected, weighted, unlabeled	similarity
15	clustering	hetero(doc+concept)	undirected, unweighted, unlabeled	semantic (bipartite graph)
16	summarization	homo(sentence)	directed/undirected, weighted/unweighted, unlabeled	similarity
17	summarization	homo(sentence)	directed/undirected, weighted, unlabeled	similarity
18	keyword extraction, summarization	homo(term or sentence)	directed/undirected, weighted/unweighted, unlabeled	co-occurrence, similarity

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

Representation", Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1332-1341, 2011.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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