텍스트 마이닝을 위한 그래프 기반 텍스트 표현 모델 활용방안

Utilization Plan of Graph-based Text Representation Model for Text Mining

저자
(Authors)

김가람, 김응모
Ga-ram Kim, Ung-mo Kim

출처
(Source)

Proceedings of KIIT Summer Conference , 2017.12, 432-433 (2 pages)

발행처
(Publisher)

한국정보기술학회
Korean Institute of Information Technology

URL

http://www.dbpia.co.kr/Article/NODE07274936

APA Style

김가람, 김응모 (2017). 텍스트 마이닝을 위한 그래프 기반 텍스트 표현 모델 활용방안. Proceedings of KIIT Summer Conference, 432-433.

텍스트 마이닝을 위한 그래프 기반 텍스트 표현 모델 활용방안

김가람(*), 김응모(**)
(*) 성균관대학교 컴퓨터공학과, kgr9937@hanmail.net
(**) 성균관대학교 소프트웨어대학, ukim@skku.edu

Utilization Plan of Graph-based Text Representation Model for Text Mining

Ga-ram Kim(*) Ung-mo Kim(**)
(*) Computer Science and Engineering, SungKyunKwan University
(**) Colleage of Software, SungKyunKwan University

요 약

본 논문에서는 다양한 연구에서 제안된 그래프 기반 텍스트 표현 모델들을 비교·분석하여 각 모델들이 실제로 어떤 곳에 사용했을 때 효과적으로 사용될 수 있을 지에 대해서 전망하고, 앞으로의 발전 방향에 대해서 논한다.

1. 서론

최근, 비/반구조화 된 텍스트 데이터에서 자연어처리 기술에 기반하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술인 텍스트 마이닝[2]을 여러 분야에서 많이 사용하고 있다. 현재까지 텍스트 마이닝에서 가장 많이 사용되고 있는 텍스트 표현 모델은 벡터공간 모델인데, 최근 들어 단어들의 의미적 관계까지 표현하기 위해 그래프를 이용한 텍스트 표현 모델에 대한 연구가 많이 진행되고 있다.[1] 본 논문에서는 그래프 기반 텍스트 표현 모델들의 실제 활용방안에 대해서 모색한다.

2. 그래프 기반 텍스트 모델의 분류

2.1 그래프 구조(format)에 따른 분류

가. 노드의 표현 방식

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 표현하는 지에 따라 동종(homogeneous) 또는 이종(heterogeneous) 표현 방식으로 난루 수 있다. 또한, 노드에 가중치를 부여할 것인지에 따라 weighted와 unweighted로 구분할 수 있다.[1]

나. 간선의 표현 방식

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]

2.2 그래프 내용(contents)에 따른 분류

가. 공기 또는 유사성 표현 모델

이 모델은 기존 연구에서 가장 많이 사용되고 있는 방식으로 단어 간의 공기 정보나 문장 간의 유사도 등을 표현한다. 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.[1]

나. 문법적 연관성 표현 모델

문법적 연관성을 표현하는 모델에서는 자연어 처리 기법을 이용하여 노드를 품사의 타입 별로 구분하고 이를 간선의 레이블로 표현함으로써 노드 간의 의존성(dependency)을 나타낼 수 있다. 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.[1]

다. 의미적 연관성 표현 모델

의미적 연관성에 대한 표현 방법은 개념을 노드로 표현하는 것이다. 대표적인 예가 문서와 개념 간의 관계를 이분 그래프로 표현한 모델이다. 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 개념 간의 관계를 트리 형태로 표현한다. 이와 같이 개념을 노드로 표현하는 방식에서는 사전에 이미 구축된 개념 집합이 존재해야 하는데 Wikipedia, WordNet 등을 이용할 수 있다. 반면에 문서와 개념 간의 이분 그래프를 이용하지만 별도의 개념집합을 사용하지 않고, 잠재적 의미 분석을 통해 선정된 문서 내의 주요 단어를 개념으로 취급한 경우도 있다.[1]

3. 활용 방안 제시

3.1 문서 분류(document classification)

문서 분류는 임의의 텍스트 문서를 이미 정해진 범주에 따라 분류하는 문제이다. 인터넷의 발전과 전산 기술의 발달에 따라 전산화된 문서의 양이 점점 더 증가하고 있고, 이에 따른 정보의 분류 문제 역시 중요한 문제로 제기되고 있다. 쉽게 표현하면 주어진 텍스트가 스포츠 분야인지, 경제 분야인지, 문화 분야인지 등을 Key Word에 따라 분류하는 것을 의미한다.[3] 따라서, 의미적 연관성 표현 모델을 사용해서 주요 범주들을 개념으로 취급하고 문서와 개념 간의 연관 관계를 이분 그래프로 표현한다면 효과적으로 수행할 수 있을 것이다.

3.2 문서 요약(document summarization)

문서 요약이란 주어진 문서로부터 특정 사용자나 작업에 적합한 축약된 형태의 문서를 생성하는 것을 말한다. 이 작업은 문서의 복잡도를 줄이면서 필요한 정보를 유지/제공할 수 있게 해준다.[4] 따라서 문서를 문법적 연관성 표현 모델로 나타낸 다음에 문서의 핵심 내용 만을 포함한 문장으로 재생성한다면 잘 요약된 문서를 얻을 수 있을 것이다.

3.3 키워드 추출(keyword extraction)

키워드 추출은 한 문서의 주제를 가장 잘 묘사하는 단어를 자동으로 착아주는 작업을 말한다.[5] 따라서 문서를 공기 또는 유사성 표현 모델로 나타낸 다음, 노드에 가중치를 부여한다면 키워드 추출을 잘 수행할 수 있을 것이다.

4. 결론

본 논문에서는 텍스트 마이닝을 효과적으로 하기 위해서 최근 주목받고 있는 그래프 기반 텍스트 모델들을 분류한 다음, 이 모델들의 구체적인 활용 방안에 대해서 논의해보았다.

참고 문헌

[1] 장재영, "텍스트 마이닝을 위한 그래프 기반 텍스트 표현 모델의 연구 동향", 한국인터넷방송통신학회 논문지, Vol.13 No.5, 2013

[2] 조성우, "BigData 시대의 기술", KT 종합기술원, 2011

[3] http://www.aistudy.co.kr/linguistics/natural/text_categorization.htm

[4] https://bi.snu.ac.kr/Courses/nlp01/d_summary.ppt

[5] https://en.wikipedia.org/wiki/Keyword_extraction