토픽모델링팀 논문 요약 - 2018312990조명재

국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법비교 - LDA와 HDP를 중심으로(박준형, 오효정, 2017)
http://koreascience.or.kr/article/JAKO201708733756400.pdf

서론
각광받고 있는 텍스트마이닝 기법인 LDA와 이를 변형한 HDP 토픽 모델링 기법을 활용하여 국내 기록 관리학 연구동향을 분석하고. 국내 기록관리학 관련 학술지 2종과 문헌정보학 관련 학술지 4종에서 1997년부터 2016년까지 발표된 기록관리학 관련 논문 1027건을 수집하고 전처리 과정을 거쳤다.
대부분의 국내 토픽모델링 연구는 대부분LDA를 사용한 것이며 HDP 토픽 모델링을 활용한 연구는 매우 드물기에 두 방식을 모두 사용해 비교해보도록 한다.

주요 시각화 도구 :
전체 토픽에 대한 막대차트 / 각 토픽 내 용어의 단어군집 / 연관 토픽에 대한 파이차트
LDAvis : Intertropic Distance Map을 제공, 각 토픽의 연관성과 prevalence(토픽 내 용어 중 전체 토픽에서 전반적으로 사용되는 용어)파악 가능
R : LDAvis package존재

연구 모델
데이터크롤링 > 전처리 (형태소분석,명사추출,벡터화) > 토픽모델링 > 결과분석

전처리과정
문서집합 생성 / 형태소 분석 / 명사 추출 / 벡터화로 진행
> 형태소 분석 과정에서 복합명사를 적절하게 처리하도록 하도록 하는 것이 중요하며
> 모든 논문에 공통으로 출연하는 어휘인 “연구” “중심” “분석”등의 어휘는 불용어로 간주
> R에서 제공하는 대표적인 형태소분석기 : RcppMeCab, KoNLP
> 어떤 형태소 분석기를 사용하느냐에 따라 상당히 다른 양상을 보일 수 있음에 주의

토픽 모델링
> 주로 Genism 라이브러리 사용, R에서는 topicmodels 라이브러리 사용
> LDA토픽 모델링은 사전에 적절한 토픽 수를 설정하는 것이 매우 중요함
> 토픽 수를 특정 범위 안에서 바꿔가면서 각 토픽 모델링을 수행한 후 적절한 토픽 수를 정하는 것이 바람직

결론 - HDA은 중복 키워드와 일반적 키보드를 배제하고, 주제별 특징을 분명하게 파악할 수 있는 키워드 반환

LDA 토픽모델링 (LSA 토픽모델링의 보완)
> 전반적으로 해당 도메인을 대표하는 주요 키워드로 빈도수에 영향을 많이 받았음
> 기록 관리학 내에 거시적으로 대표되는 주제들 도출에 유용

HDP 토픽모델링(추가적인 텍스트마이닝 기법)
> 토픽별 특징을 파악할 수 있는 특수한 키워드를 많이 도출할 수 있음
> 세부 주제별 미시적 핵심 키워드 도출에 효과적

논문에서는 LDA와 HDA를 같은 전처리 과정을 거친 기록관리학 관련 문서 집합을 대상으로 수행하여 그 결과가 상이하게 나타남을 확인함. HDP의 경우 단순히 문서 집합에서 빈도수만 높고 일반적인 키워드보다 토픽의 주제와 특징을 파악할 수 있는 특수한 키워드가 높은 순위에 있는 경향을 보임. LDA와 비교했을 때 상대적으로 세부적인 주제를 파악할 수 있는 키워드가 많기 때문에, LDA 토픽모델링 보다 더욱 세부적인 연구동향 분석에 유용하다, 반면에 LDA는 해당 분야에서 공통적으로 많이 다루고 있는 주제 키워드 분석에 유리하여 거시적인 연구동향 파악에 적합한 방법으로 보임.

+자세한 이론적 배경은 직접 정리하기엔 너무 복잡해서 따로 워드로 만들어 보았습니다.

토픽모델링팀 논문 요약 - 2018312990조명재

문헌정보학 분야 핵심 학술지의 가중 주제-방법 네트워크 분석(이기헌, 정효정, 송민, 2015)
https://kslis.accesson.kr/assets/pdf/9364/journal-49-3-457.pdf

서론/연구목적
- 1990년부터 2014년까지 25년간 국외 유수 문헌정보학 학술자들에 게제된 논문을 대상으로 연구주제와 연구방법을 구분하여 학술지의 연구 누적 현황을 분석, 텍스트 마이닝과 LDA 기법을 활용함.
- 연구 주제와 연구 방법을 구분하고 확률적 토픽 모델링을 적용하여 토픽을 추출한 다음, 추출한 토픽을 연구주제/연구방법으로 구분한 뒤 동시출현 토픽을 기초로 가중 주제-방법 네트워크를 생성한다. 이 과정을 통해 주제와 방법의 관계에서 두드러지는 주제와 방법을 확인하고, 괄목할만한 주제-방법의 관계들을 살펴본다.

- 학술데이터의 양이 늘어나면서 인간이 전체 데이터의 내용을 분석하는 것이 어려워 정량적 분석이 필요함.
- 출연단어 기반의 네트워크 분석은 출연 단어의 수가 많을수록 관계가 기하급수적으로 증가하기 때문에 몇 십만개의 단어들과 그 이상의 연결관계를 관찰하고 해석하는건 힘들다
  - 군집화와 같은 텍스트 마이닝 기법을 적용해 이해할 수 있을 만한 수의 군집들로 단어를 분류
- 개념 추출과 이에 대한 해석이 여전히 난해함
  - 토픽 모델링을 통하여 면밀하게 보완분석, 가중주제-방법 네트워크 분석 모델 제안

연구방법
- 대상 문헌들 내의 어휘들을 군집화 / 개념화한 뒤, 인간이 이해할 수 있을만한 수의 주제들과 방법들을 도출
- 이 **주제와 방법들을 점으로 갖는 네트워크를 구축, 이때 이 연결선은 특정 주제에 관련된 어휘와 특정 방법에 관련된 어휘가 같은 문허에서 동시에 출현하는 빈도를 의미, 한 문헌에서도 여러번 등장 할 수록 가중된 값을 갖도록 함**

데이터 수집
- 선도 연구는 타 연구에 대한 영향력 정도로 정의, 84개의 학술지 중 가장 영향력이 높은 20개의 학술지를 선정
- 1990년부터 2014년 사이의 논문들 20344개를 선정, 학술지 당 평균 논문 수 1051개
-

데이터 전처리
- Stanford NLP 형태소 분석기 사용

주제와 방법 추출
- LDA기법을 활용하여 토픽을 설정.
- 문헌 구성 단어 - 특정 토픽에 할당되어야 함
- 토픽에 할당되는 단어들을 이산 확률 분포인 다항분포로 표현하여 문헌을 입력받을 때 마다 통계적으로 추론. 추론이 끝나면 모든 문헌에 속한 모든 단어들이 각각 사용자가 지정한 수 만큼의 “토픽”에 분류된다.
- 토픽은 모든 문헌에서 발견된 단어들의 확률 분포로, 동일한 단어가 모든 토픽에 있어도 출현 확률과 다른 단어의 출현 확률들로 토픽의 유일한 특성을 갖게 됨
- 개념(토피)의 이름 부여와 주제/방법 구분은 해당 분야 전문가들의 도움으로 수행, 50개의 최빈 단어들과 해당 개념이 차지하는 비중이 높은 논문들에 대한 서지사항을 제공하여 개념을 명명하고 주제나 방법으로 분류
- 참고한 분류체계 : JITA Classification System of Library and Information Science

가중 주제-방법 네트워크 분석
네트워크의 노드들인 주제와 방법을 분석, 주제 연결 중심성(다양한 방법들과 어느정도 연결되어 있는지 표현) / 방법 연결 중심성(다양한 주제들과 어느정도 연결되어 있는지 표현) / 가중 연결 중심성 / 매개중심성 등을 중심적으로 조명

연구결과 : 논문 본문 참고(토픽별 핵심어 추출, 연결중심성, 가중연결중심성, 매개중심성 등등)