키워드 요약의 세 가지 방법론 비교

강종렬, 남지성, 박지나, 김웅섭
동국대학교 정보통신공학과
e-mail: jongreulk@dongguk.edu, njs51@dongguk.edu, zxzxzi@naver.com, woongsup@dongguk.edu

Compare Three Method for Keyword Summary

Jong-Reul Kang, Ji-Seong Nam, Gi-na Park, Woongsup Kim,
Dept of Information Communication Engineering, Dongguk University, Seoul, Republic of Korea

요 약

본 논문은 정확한 연관검색어를 보여주지 못하는 기존의 검색에서 벗어나기 위해 이미지와 PDF에서 텍스트를 추출하고 키워드 요약하는 방법을 사용하였다. 텍스트를 키워드로 요약하는 알고리즘으로는 TextRank, LSA, MMR을 사용하였고, 세 가지 방법으로 키워드를 요약하고 키워드 요약 결과와 Query의 코사인 유사도를 이용하여 추출한 문서와 Query와의 연관성을 확인하여 세 가지 알고리즘을 비교하였다.

1. 서론

기존의 키워드 기반 검색은 해당 단어를 입력하면 단어가 포함된 문서에 같이 등장하는 단어 중 빈도수가 높은 단어를 보여준다. 이로 인해 불용어나 공통 단어가 연관검색어로 나타나 사용자의 의도와 일치하지 않는다. 본 논문에서는 이를 해결하기 위해 이미지나 PDF와 같은 첨부파일에서 텍스트를 추출하고 입력문서에 통합함으로써 연관검색의 정확성을 높였다. 또한, 세 개의 알고리즘을 이용하여 문서를 여러 개의 키워드로 요약하여 불용어와 공통 단어를 기존보다 많이 필터링해 정보의 중복성을 줄였다. 각각의 방법으로 요약된 키워드와 Query와의 코사인 유사도를 이용하여 가장 효율적인 알고리즘을 찾고자 한다.

2. 관련 연구

2.1 TextRank

PageRank[1]는 구글에서 제안한 알고리즘이며 기본적으로 중요한 사이트는 다른 많은 사이트로부터 링크를 받는다는 점에 착안하여 문서 A에서 가지고 있는 B의 링크가 B 사이트를 추천하는 한 표로 해석하여 이를 기준으로 중요도를 평가한다. TextRank[2] 알고리즘은 PageRank의 중요도가 높은 웹사이트는 다른 많은 사이트로부터 링크를 받는다는 점에서 사이트를 문장 혹은 단어로 생각하고 문서 내의 문장, 단어의 중요도를 계산하는 알고리즘이며 그 식은 (1)과 같다.

$$TR(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} TR(V_j) \quad \cdots (1)$$

여기서 $TR(V_i)$는 문장 또는 단어 V_i에 대한 TextRank 값을 의미하고 w_{ij}는 문장 또는 단어 i와 j 사이의 가중치, d는 dampingfactor로써 PageRank에서 웹 서핑을 하는 사람이 해당 페이지를 만족하지 못하고 다른 페이지로 이동하는 확률로써, TextRank에서도 그 값을 그대로 0.85로 사용한다.

본 연구에서는 불용어 설정, 제거 후 $TR(V_i)$를 계산 한 뒤 높은 순으로 문서들을 키워드로 요약하였다.

2.2 Latent Semantic Analysis

LSI(Latent Semantic Index) 라고도 하는 LSA (Latent Semantic Index)[3]는 Bag of Word(Bow) 모델을 사용하여 문서 용어 행렬을 생성한다. 행은 용어를 나타내고 열은 문서를 나타낸다. LSA는 특이 값 분해(SVD)를 사용하여 문서 용어 행렬에 대해 행렬 분해를 수행하여 잠재적인 주제를 학습한다. LSA는 일반적으로 치수 축소 또는 노이즈 감소 기술로 사용된다. LSA 알고리즘은 이해하고 구현하기 쉬운 가장 간단한 방법이다. 또한 벡터 공간 모델에 비해 더 나은 결과를 제공한다. 문서 용어 행렬 분해만 포함하기 때문에 다른 사용 가능한 알고리즘에 비해 빠르다.

본 연구에서는 LSI를 이용하여 문서를 하나의 주제, 10개의 키워드로 요약하였다.

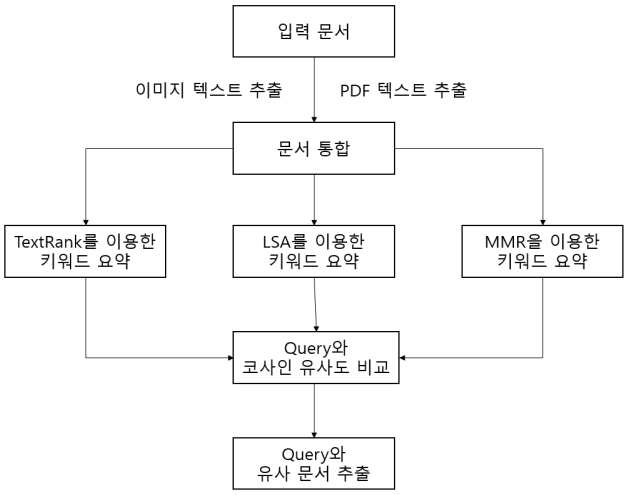
2.3 MMR

일반적인 MMR(Maximal Marginal Relevance)[4] 알고리즘의 수식은 (2)와 같다.

$$MMR = \arg\max [\lambda Sim_1(D_i, Q) - (1-\lambda) \max Sim_2(D_i, D_j)] \cdots (2)$$

λ가 작아질수록 정보의 중복성에 관계없는 결과가 출력된다.
MMR은 고려중인 항목과 이미 선택된 항목 사이의 불일치를 정량화하기 위한 척도다. MMR이 높을수록 고려 된 항목이 Query와 관련이 있으며 이전에 선택한 항목과의 유사성이 거의 없음을 의미한다. 따라서 가장 빈도 수가 높은 단어 n개를 추출 후 Query에 저장해 그와 유사한 문장을 찾은 후, 다음의 문장을 찾을 때엔 Query와 유사하지만 이전 문장과의 유사도가 낮은 문장을 추출하는 것을 목표로 한다.
본 연구에서는 비슷한 내용의 문장이 반복되는 것을 피하고 정보의 중복성을 제거하기 위해 사용되었으며 λ를 0.5로 지정하였다.

3. 연구 방법



(그림 1) 구성도

본 논문에서의 시스템 형태는 그림 1과 같다. [5]와 같이 입력 문서에서 GOOGLE CLOUD로 부터 이미지, PDF-Miner를 통해 PDF를 텍스트로 추출하였다. 이를 이용해 3가지 방법으로 키워드를 요약하고 Query와의 코사인유사도를 2가지 방법으로 비교하였다.
3가지 방법에서 각각 Query와 유사한 문서를 선택하는데 코사인 유사도를 이용하였다. 그 식은 (3)와 같다.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad \cdots (3)$$

위 식을 이용하여 세 가지 방법으로 유사도가 높은 10개의 문서를 추출하였고 Query와의 실제 연관성을 비교하였다.

4. 연구 결과

본 논문에서는 실험에 동국대학교 웹페이지의 일반, 학사, 국제 부분의 공지사항 450개를 이용하였다. 이를 이용해 3가지 방법으로 공지사항들을 분석하여 10개의 키워드로 요약하고 Query와 각각의 키워드간의 코사인 유사도 평균, 키워드 전체의 벡터 평균과 Query와의 코사인 유사도를 비교하였다.
'기숙사' Query와 3가지 키워드 요약방법들과의 코사인 유사도는 표 1과 같으며 해당 수치는 소수점 넷째 자리에서 반올림하여 구했다.

<표 1> 세 가지 알고리즘과 '기숙사' Query와의 코사인 유사도

	유사도 평균	벡터 평균
MMR	2.18068/10 = 0.219	4.95307/10 = 0.495
LSA	2.24611/10 = 0.222	4.88746/10 = 0.489
TextRank	2.15409/10 = 0.215	4.72717/10 = 0.473

결과를 보면, 벡터 평균방법이 유사도 평균 방법보다 대략 0.25 정도 코사인 유사도가 높은 것을 볼 수 있다.
또한, 코사인 유사도를 이용하여 '기숙사' Query와 유사한 문서를 3개의 방법에서 각각 10개 선택하였고 실제 '기숙사' Query와의 연관성을 제목과 본문을 통해 확인하였다. 그 결과는 표 2와 같다.

<표 2> 세 가지 알고리즘으로 뽑은 문서와 '기숙사' Query와의 연관성 비교

	유사도 평균 문서	벡터 평균 문서
MMR	6/10 = 0.6	6/10 = 0.6
LSA	5/10 = 0.5	7/10 = 0.7
TextRank	3/10 = 0.3	6/10 = 0.6

표 1과 같이 표 2에서도 유사도 평균 방법보다 벡터 평균 방법이 Query와의 관련성이 높은 것을 볼 수 있다.

5. 결론 및 향후 연구 방향

본 논문에서는 TextRank를 포함하여 LSA, MMR 세 가지 알고리즘으로 키워드를 요약하였다. 요약된 키워드와 Query의 코사인 유사도를 이용하여 10개의 문서를 추출하였고 Query와의 실제 연관성을 확인해보았다. MMR, LSA 알고리즘을 이용하여 정확한 유사 문서나 문장을 찾거나 코사인 유사도 가 아닌 다른 방법론에 대한 추후 연구가 필요하다.

참고문헌

[1] S.Brin and L.Page, "The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems, vol.30, no.1-7, pp.107-117, 1998.

[2] Rada Mihalcea, Paul Tarau, “TextRank: Bringing Order into Texts”, Proceedings of the European Conference on Artificial intelligence (ECAI 2004), Valencia, Spain, August 2004.

[3] Thomas K Landauer, Peter W. Foltz & Darrell Laham (1998) An introduction to latent semantic analysis, Discourse Processes, 25:2-3, 259-284

[4] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR 1998, pages 335 - 336, 1998

[5] 강종렬, 김웅섭, “문서요약 및 키워드 추출을 통한 태그 클라우드 연관 검색어 제시법”, 한국인터넷정보학회 춘계학술발표대회, 2019