다양한 차원 축소 기법을 적용한 문서 군집화 성능 비교

조 희 련
국민대학교 소프트웨어융합대학
e-mail : heeryon@kookmin.ac.kr

Comparison of Document Clustering Performance Using Various Dimension Reduction Methods

Heeryon Cho
College of Computer Science, Kookmin University

요 약

문서 군집화 성능을 높이기 위한 한 방법으로 차원 축소를 적용한 문서 벡터로 군집화를 실시하는 방법이 있다. 본 발표에서는 특이값 분해(SVD), 커널 주성분 분석(Kernel PCA), Doc2Vec 등의 차원 축소 기법을, K-평균 군집화(K-means clustering), 계층적 병합 군집화(hierarchical agglomerative clustering), 스펙트럼 군집화(spectral clustering)에 적용하고, 그 성능을 비교해 본다.

1. 서론

문서 군집화는 많은 양의 문서를 문서에 대한 분류 레이블 없이 자동으로 유사 문서들로 묶어 주는 문서 처리 기법이다. 인터넷의 등장으로 누구든지 자유롭게 인터넷에 의견을 개진하게 되면서, 문서 또는 텍스트 형태의 다양하고 많은 양의 의견들을 자동으로 구분하고 조망하려는 요구가 나날이 늘고 있다. 동시에 이러한 요구에 부응하는 기술로 문서 군집화 기술이 각광을 받고 있다.

문서 군집화는 전통적으로 문서 내 단어 출현 횟수 등을 정의한 문서 벡터를 가지고 군집화를 실시한다. 이러한 문서 벡터 생성에는 TF-IDF 알고리즘[1]이 주로 사용돼 왔는데, TF-IDF는 하나의 단어를 하나의 벡터 요소, 즉, 다차원 공간 상의 하나의 축으로 정의하기 때문에 단어들 간의 관계를 표현하지 못한다는 단점이 있다. 이러한 단점을 극복하기 위해 특이값 분해(singular value decomposition: SVD)를 이용한 문서 벡터의 차원 축소가 제안되었으며, 자연어처리 분야에서는 SVD를 흔히 잠재의미분석(latent semantic analysis: LSA)[2]이라 부른다. LSA로 차원 축소된 문서 벡터는 단어들 간의 관계 중에서도 동의어 및 동음이의어를 파악하는데 유용하다. SVD 나 주성분분석(principal component analysis: PCA)은 문서 벡터를 저차원 공간으로 선형 변환하는데, 고차원의 내적(inner product) 계산을 통해 비선형 변환을 실현하는 차원 축소 기법으로 Kernel PCA[5]가 제안되었다.

최근 들어 심층 신경망 학습(deep neural network learning)이 문서 처리에도 적용되면서, 고차원의 희소한 단어 벡터(sparse word vector, 흔히 one-hot encoding으로 정의)를 저차원 밀집 벡터(lower rank dense vector)로 표현하는 단어 임베딩(word embedding) 기법[3]이 제안되었고, 문서 벡터를 통째로 저차원 공간에 임베딩하는 Doc2Vec[4]도 제안되었다.

차원 축소는 주어진 데이터의 노이즈를 제거하고 특징을 추출하는 효과가 있어 다양한 데이터 처리에 활용되고 있는데, 본 발표에서는 문서 군집화를 위한 문서 벡터의 차원 축소에 초점을 맞춰 다양한 차원 축소 기법과 다양한 문서 군집화 알고리즘의 결합이 어떤 결과를 가져오는지를 비교하고자 한다. 이를 위해 문서 군집화에 자주 쓰이는 벤치마크 데이터셋에 SVD, Kernel PCA, Doc2Vec 의 차원 축소 기법과 K-평균 군집화, 계층적 병합 군집화, 스펙트럼 군집화를 결합, 적용하여 이들의 성능을 비교한다.

2. 차원 축소 기법

SVD: N개의 문서를 행으로 삼고 W개의 특징 단어를 열로 삼는 $N \times W$행렬로 구성된 문서 집합이 주어졌을 때, 특징 단어의 열이 k ($k \ll W$) 차원으로 축소된 문서 집합은 다음의 행렬 분해로 구할 수 있다.

$$X \approx X_k = U_k \Sigma_k V_k^T$$

여기서 Σ_k는 대각행렬이며, $U_k \Sigma_k$를 통해 차원 축소된 문서 벡터의 집합을 구할 수 있다.

Kernel PCA: W차원 공간의 문서에 어떤 비선형 함수 $\phi(x) \in \mathbb{R}^k (k > W)$로 문서 집합 X를 k차원 공간으로 변형한 뒤 문서 집합 $\{\phi(x_n)\}_{n=1}^N$에 대하여 k차원 공간 상에서 PCA를 적용한다.

Doc2Vec: 단어 벡터와 단락 벡터(paragraph vector)를 결합하여 문장 전체의 벡터를 구하며, 문장 속 단어를 예측하면서 동시에 문장의 분산 표현(distributed representation)을 예측하는 방식으로 문서 벡터를 구한다.

3. 군집화 기법

K-평균 군집화: 같은 군집에 속한 데이터는 서로 가까이에 놓여있다는 가정 하에 군집의 중심에 있는 데이터(centroid)와, 군집에 속한 나머지 데이터들 간의 거리가 최소화되도록 군집을 구성한다.

계층적 병합 군집화: 각 데이터가 저마다 하나의 군집을 이루는 것에서 출발하여 가까운 데이터들을 병합하여 차차 군집의 개수를 줄임과 동시에 군집의 크기를 키워나가며 군집화한다. Ward, complete linkage, average linkage 등의 알고리즘이 있다.

스펙트럼 군집화: 문서 집합으로 그래프를 생성하여 그래프의 연결 성분 분석을 통해 군집화를 진행한다. 그래프에서 인접행렬과 차수행렬을 구하여 라플라시안 행렬을 구하고, 고유값 분해를 통해 고유값을 구한 뒤, 고유값이 작은 고유벡터의 일부를 선택한 뒤 K-평균 군집화 등의 군집화 알고리즘을 적용한다.

4. 실험

데이터셋: 문서 군집화 평가 실험을 위해 미국 4 개 대학의 영문 홈페이지를 수집한 'The 4 Universities Data Set (http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/)을 벤치마크 데이터셋으로 사용했다. 데이터셋의 구성은 <표 1>과 같다. 문서 전처리로 각 HTML 문서의 HTML 태그와 영어 불용어를 제거했고, 남은 단어들로 단어를 벡터 요소로 하는 문서 벡터를 만들었다.

<표 1> 실험에 사용한 미 대학 홈페이지 문서의 구성

문서 카테고리	문서 개수
Course	244
Department	4
Faculty	153
Project	86
Staff	46
Student	558
Total	1,091

실험환경: TF-IDF 로 21,597 개의 단어로 구성되는 기본적인 문서 벡터를 구축하고 이 문서 벡터에 SVD, Kernel PCA, Doc2Vec 의 차원 축소 기법을 적용하여 차원이 300 인 문서 벡터를 구축하였다. 군집화 알고리즘으로는 K-평균 군집화, 계층적 병합 군집화(Ward, complete linkage, average linkage 의 세 군집화 기법), 스펙트럼 군집화를, 군집 개수를 문서 카테고리 개수와 동일한 6 개로 설정하여 적용했다. 문서 벡터 생성과 차원 축소, 군집화에는 공개 기계학습 소프트웨어인 scikit-learn (http://scikit-learn.org)을, Doc2Vec 구축에는 Genism (https://radimrehurek.com/gensim/)을 사용했다.

평가척도: 군집화 성능 평가에 가장 많이 사용되는 normalized mutual information (NMI)과 adjusted rand index (ARI)를 평가 척도로 사용했다. NMI 의 경우 평가값이 1 에 가까우면 군집화 성능이 좋은 것으로 해석되고, 0 에 가까우면 군집화 성능이 나쁜 것으로 해석된다. ARI 의 경우 평가값이 양수이면 군집화 성능이 괜찮고, 음수이면 군집화 성능이 나쁘다.

<표 2> Normalized mutual information 을 이용한 평가

NMI	TF-IDF	SVD	Kernel PCA	Doc2 Vec
K-Means	0.344	0.348	**0.360**	0.274
Ward	0.225	0.242	0.266	0.249
Complete	0.155	0.098	0.066	0.045
Average	*0.010*	0.055	0.055	0.036
Spectral	0.168	0.122	0.081	0.078

<표 3> Adjusted rand index 를 이용한 평가

ARI	TF-IDF	SVD	Kernel PCA	Doc2 Vec
K-Means	0.282	0.301	**0.325**	0.304
Ward	0.108	0.171	0.274	0.323
Complete	0.089	0.044	0.025	0.017
Average	*0.002*	0.009	0.009	0.004
Spectral	0.085	0.034	0.034	0.004

5. 결과 및 결론

<표 2>와 <표 3>에 NMI 와 ARI 로 평가한 실험 결과를 제시한다. 두 평가척도 모두 K-평균 군집화와 Kernel PCA 의 결합을 가장 좋은 군집화 결과로 평가했다. 반면에 계층적 병합 군집화의 average linkage 알고리즘과 TF-IDF 의 결합을 가장 나쁜 군집화 결과로 평가했다. Doc2Vec 을 활용한 차원 축소는 데이터가 적은 탓인지 기대했던 것보다 낮은 군집화 성능을 나타냈다. 또 스펙트럼 군집화도 기대에 못 미치는 군집화 성능을 보였다.

감사의글

이 논문은 2018 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2017R1A2B4011015).

참고문헌

[1] Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation. 28: pp. 11–21.

[2] Susan T. Dumais (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology. 38: pp. 188–230.

[3] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeff (2013). "Distributed representations of words and phrases and their compositionality". Advances in Neural Information Processing Systems.

[4] Quoc Le and Tomas Mikolov. (2014). "Distributed representations of sentences and documents". In Proc. of the 31st International Conference on Machine Learning (ICML'14).

[5] Schölkopf, Bernhard (1998). "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". Neural Computation. 10: pp. 1299–1319.