가중치를 이용한 효과적인 항공 단문 군집 방법

김주영*, 이지민*, 안순홍*, 이훈석*
*아시아나 IDT
e-mail : kimjy6@asianaidt.com

DOCST: Document frequency Oriented Clustering for Short Texts

Jooyoung Kim*, Jimin Lee*, Soonhong An*, Hoonsuk Lee*
*AsianaIDT

요 약

비정형 데이터의 대표적인 형태 중 하나인 텍스트 데이터 기계학습은 다양한 산업군에서 활용되고 있다. NOTAM 은 하루에 수 천개씩 생성되는 항공전문으로써 현재는 사람의 수작업으로 분석하고 있다. 기계학습을 통해 업무 효율성을 기대할 수 있는 반면, 축약어가 혼재된 단문이라는 데이터의 특성상 일반적인 분석에 어려움이 있다. 본 연구에서는, 데이터의 크기가 크지 않고, 축약어가 혼재되어 있으며, 문장의 길이가 매우 짧은 문서들을 군집화하는 방법을 제안한다. 주제를 기준으로 문서를 분류하는 LDA 와, 단어를 k 차원의 벡터공간에 표현하는 Word2Vec 를 활용하여 잡음이 포함된 단문 데이터에서도 효율적으로 문서를 군집화 할 수 있다.

1 장. 서론

사진, 음성 등과 더불어 비정형 데이터로서 많이 쓰이는 텍스트 데이터 기계학습은 다양한 분야에서 널리 쓰이고 있는 방법이다. 항공분야에서도 업무 효율성 극대화를 통한 방법으로서 기계학습방법을 적용하고자 노력하고 있다. NOTAM(노탐, Notice to Airmen)은 항공관련 시설 등에 변경 및 이상요인이 생겼다는 사실을 포함하고 있는 전문으로서, 축약어가 혼재된 단문 데이터이다. 하루에 수 천개씩 생성되는 NOTAM 을 운항관리사가 일일이 확인하는 현재의 업무 프로세스에 기계학습을 접목시킬 경우 업무 효율성이 상승할 것으로 기대된다. 하지만, 축약어가 혼재된 단문 데이터인 NOTAM 은 희박함(Sparsity) 문제를 갖고 있어서 일반적인 분석이 힘들다는 단점이 있다.

본 연구에서는, 짧은 문장에 대해서 가중치를 활용한 주제와 의미 기반의 효율적 군집 방법을 제안하고자 한다. 본 연구에서 제안하는 방법은 크게 3 가지 특징으로 이루어져있다. 먼저, 자료의 희박함 문제를 해결하고자 *LDA*(Latent Dirichlet Allocation)을 사용한다. *LDA*는 주제분포를 통해 1 차적으로 문서를 군집해주는 역할을 수행한다. 분류된 군집에 사용된 단어 수는 문서 전체의 총 단어 수 보다 줄어들기 때문에 자료의 희박함 문제를 일정부분 해결해줄 수 있다. 다음으로, 의미기반의 군집을 위해 *Word2Vec* 를 이용하였다. 마지막으로, *Word2Vec* 를 통해 나온 단어들의 벡터에 가중치를 부여함으로써 서로 다른 의미를 지닌 단어들이 벡터공간에서 분리 될 수 있도록 하였다.

본 논문은 5 장으로 구성되어 있다. 1 장 서론에 이어 2 장에서는 본 연구에 사용된 방법들을 소개한다. 3 장에서는 제안하는 알고리즘 설명을 포함하고 있으며, 항공 관련분야의 텍스트 데이터인 NOTAM 의 분석 결과를 4 장에 담았다. 마지막으로, 5 장에서는 본 연구의 요약과 함께 향후 연구에 대한 내용을 담았다.

2 장. 관련 연구

2-1. *LDA*

확률적 주제 모형(Probabilistic Topic Model) 중 하나인 *LDA*(Latent Dirichlet Allocation)는 각 문서에 어떤 주제들이 존재하는지를 확률적으로 표현해주는 모형이다.

초기에 Hofmann(1999)는 Probabilistic Latent Semantic Indexing (pLSI)이라는 확률적 주제 모형을 제안하였다. 이 모형은 한 문서에 존재하는 주제의 분포에 대한 가정이 없어서 여러 가지 한계점을 갖고 있었다. Blei(2003)은 이 주제의 분포에 디리클레 사전확률(Dirichlet prior)를 적용하여 *LDA* 모형을 제안하였다.

2-2. *Word2Vec*

Word2Vec(Mikorov, 2013)는 2002 년 Bengio 에 의해 개발된 NNLM(Neural Net Language Model)[2]을 개선한 방법으로, NNLM 의 Hidden Layer 을 제거한 후, Hierarchical Softmax 등의 알고리즘을 추가하여 기존의 방법에 비해 몇 배 이상 빠른 학습을 가능하도록 만든 모형이다.

Word2Vec 는 CBOW(Continuous Bag-of-word)와 Skip-gram 이라는 2 가지 네트워크 모형을 포함하고 있다. 대용량의 데이터에서는 CBOW 보다 Skip-gram 이

더 좋은 성능을 보인다고 알려져 있다[4]. 하지만, 짧은 문장을 학습할 경우에는 CBOW 의 성능이 좀 더 좋다는 연구결과도 있다[6].

2-3. 단어 좌표를 이용한 문서 좌표화

단어의 좌표를 활용하여 문서의 좌표를 나타내는 방법은 다음과 같다.

$$\text{Mean: } \mathrm{M} = \frac{1}{n_m}\sum_{j=1}^{n_m} w_j$$

n_m: 문장 M 에 포함된 단어 개수

w_j: j 번째 단어 벡터

위의 방법은, 하나의 문서를 구성하는 단어들의 좌표를 단순히 평균하여 문서를 하나의 벡터로 만드는 방법이다. Cedric De Boom[7]는 idf (inverse document frequency)를 이용하여 특정 단어들에게 가중치가 부여되어 만들어진 문서 벡터를 제안하였다.

$$\text{CDB: } C = \frac{1}{n_c}\sum_{j=1}^{n_c} idf_j \cdot w_j \ , \qquad idf_j = \log\frac{N}{df_j}$$

idf를 이용한 가중문서벡터 C는 idf가 큰 단어들의 영향을 받게 되어 벡터공간에서 위치가 조정된 문서 벡터가 된다.

3 장. 알고리즘

3-1. SDF

idf 는 tf와 반대로 단어의 희소성을 표현하는 지표이다. idf가 높다는 뜻은, 그 해당 단어는 특정 문서에서만 나온다는 뜻이다. 본 연구에서는 idf의 특징을 이용하여 다음과 같은 가중치를 제안한다.

$$Softmax\ DF: SDF_j = \frac{\exp(\Delta_j)}{\sum_j \exp(\Delta_j)}, \qquad \Delta_j = \max_i(idf_i) - idf_j$$

위 가중치는 idf와 idf의 최대값 차이에 $Softmax$ 함수를 사용하여 나온 결과값이다. Akaike weight[3]와 비슷한 형식이지만, 최대값을 뺐다는 점에서 차이가 있다. 가중치의 분자를 정리하면 다음과 같다

$$\exp(\Delta_j) = \exp(max(idf) - idf_j)$$

$$= \frac{exp\,(max(idf))}{\exp(idf_j)} = \frac{\exp\left(\max_i\left(\log\frac{N}{df_i}\right)\right)}{\exp\left(\log\frac{N}{df_j}\right)}$$

$$= \frac{\frac{N}{\min_i(df_i)}}{\frac{N}{df_j}} = \frac{df_j}{\min_i(df_i)}$$

가중치의 분모는 상수로 고정되어 있기 때문에 가중치는 아래의 식으로 정리 할 수 있다.

$$SDF_j \propto \frac{df_j}{\min_i(df_i)}$$

가중치 SDF_j는 자주 등장하는 단어의 경우에는 큰 가중치를, 자주 등장하지 않는 단어들에겐 작은 가중치를 부여한다. 따라서, k차원의 좌표평면에 위치한 단어들의 위치가 재조정된다.

3-2. *DOCST*

본 연구의 알고리즘은 LDA 로 나눈 군집들을 각각 Word2Vec 로 학습한 후 K-Means 로 군집하는 방법이다.

DOCST: Document frequency Oriented Clustering for Short Texts
Step 1: LDA를 통해 전체 문서 D를 t개의 주제군집 L로 나눈다. $D = \{L_1, \dots, L_t\}$
Step 2: $i = 1, \dots, t$ 2-1. i 번째 주제군집 L_i 의 문서-단어 결합 행렬 (Document-term incidence matrix, DTM) DTM_i을 구한다. (DTM의 값은 단어가 존재하면 1, 아니면 0 이다) 2-2. L_i를 Word2Vec에 학습하여 k차원의 벡터공간 $W2V_i$를 구한다. $W2V_i = \{W2V_{i_1}, \dots, W2V_{i_{n_i}}\}$ 2-3. L_i 에 있는 모든 단어들의 가중치 SDF_i를 구한다. $SDF_i = \{SDF_{i_1}, \dots, SDF_{i_{n_i}}\}$ 2-4. $W2V_i$에 가중치 SDF_i를 곱한 가중벡터공간 $W2V_i^{SDF}$을 구한다. 2-5. DTM_i과 $W2V_i^{SDF}$의 내적을 통해 $D2V_i$를 구한다. $D2V_i$ (k x p) = DTM_i (k x n_i) × $W2V_i^{SDF}$ (n_i x p)
Step 3. $Kmeans$를 통해 $D2V_i$를 k개의 하위군집 $S_{i,j}$로 나눈다. $S_i = \{S_{i_1}, \dots, S_{i,j}, \dots, S_{i_k}\}$
결과: $t \times k$ 개의 군집 $\{S_{1,1}, \dots, S_{t,k}\}$

<표 1> *DOCST*

표현	설명
t	주제의 수.
i	주제군집 번호. $i = 1, \dots, t$
D	전체 문서
L_i	i번째 주제군집
n_i	L_i의 총 단어 개수
DTM_i	L_i의 문서-단어 결합 행렬
$W2V_i$	L_i 단어의 k차원 벡터공간
SDF_i	L_i의 단어 가중치
$W2V_i^{SDF}$	L_i 단어의 k차원 가중벡터공간
$D2V_i$	L_i 문서의 k차원 벡터공간
k	하위군집의 수
j	하위군집 번호 $j = 1, \dots, k$
S_i	L_i의 하위군집들
$S_{i,j}$	L_i의 j번째 하위군집

<표 2> *DOCST*에 쓰인 기호

4 장. 실험

NOTAM 은 항공로, 공항시설, 항행안전시설의 설치 또는 변경, 업무처리절차의 변경, 위험요인이 생겼다는 사실을 포함하고 있는 전문으로서, “신속한 전송,

해석상의 오류를 방지하기 위해 국제민간항공기구(ICAO)가 제정한 부호, 형식" 으로 육·해·공군·미군·국토해양부 등 각 기관의 독자적인 통신망으로 수발되고 있다. [9]

ID	NOTAM (문서의 수: 4000, 단어의 수: 6341)
1	RESTRICTED AREA ACT UUR271
2	W532N ACT
3	THE SEGMENT NUKTI 40KM WEST OF NUKTI OF ATS RTE B215 CLSD AT 9500M AND BELOW
4	TWY G CLSD DUE TO GWANGJU/JEONAM PRELIMINARY OF SPACE CHALLENGE 2016
5	ATS RTE Y655 CLSD DUE TO MISSILE LAUNCHING BY DPRK
…	….
3999	WILLIAMTOWN WEST AIRSPACE R559A … RESPONSIBILITY TO CK AND MNT STS
4000	DAIL HR OF OPS OF SIEM REAP INTL AP IS NOW FM 2300 TO 1900

<표 3> 실험에 사용한 NOTAM 데이터 요약 및 예제

NOTAM 은 축약어가 혼재되어 있으며, 문장의 길이가 매우 다르다는 특징이 있어서 자료의 희박함이 매우 높기 때문에 일반적인 군집방법으로는 문서를 분류하기 어렵다.

본 장에서는 NOTAM 을 이용하여 3 가지 실험을 진행하였다. 첫 번째는 본 연구방법으로 나누어진 군집의 코사인 유사도를 2-4 에서 설명한 *Mean* 과 *CDB* 의 결과와 비교하였다. 두 번째는, 본 논문의 방법과 다른 문서 군집방법 코사인 유사도 비교해보았다. 마지막 실험은, *SDF* 유무에 따라서 실제로 데이터가 어떻게 나누어지는지 알아보고자 한다.

실험에서 사용한 값으로서 *LDA* 의 디리클레 분포 모수 α, β 는 각각 0.1, 1.0 으로 설정하였다. *Word2Vec* 에서는 벡터공간의 차원은 100, window 크기는 1 로 하였다.

4-1. 첫 번째 실험

첫 번째는, 주제군집 수 t와 하위군집 수 k의 변화에 따른 3 가지 방법(*Mean*, *CDB*, *DOCST*)의 코사인 유사도를 비교해보았다. 실험 4-1 에서 확인한 군집 크기의 변화의 결과로서 각 군집의 유사도가 어떠한지 알아보고자 한다. 실험결과는 20 번 반복한 결과의 평균이다.

총 군집 수	(t, k)	*Mean*	*CDB*	*DOCST*
32	(2, 16)	0.4318	0.3660	0.4914
	(4, 8)	0.4118	0.3645	0.4765
	(8, 4)	0.4104	0.3803	0.4671
	(16, 2)	0.4100	0.3816	0.4446
48	(2, 24)	0.4878	0.4335	0.5408
	(4, 12)	0.4765	0.4127	0.5267
	(8, 8)	0.4670	0.4223	0.5174
	(16, 3)	0.4584	0.4228	0.5019
64	(2, 32)	0.5237	0.4760	0.5669
	(4, 16)	0.5127	0.4595	0.5593
	(8, 6)	0.5035	0.4530	0.5503
	(16, 4)	0.4961	0.4564	0.5377

<표 4> 모든 군집의 평균 코사인 유사도

[표 4]의 결과를 통해 본 논문의 방법이 다른 두 가지 방법인 *mean* 과 *idf* 에 비해 좋은 성능을 보이고 있음을 확인 할 수 있다. 전체적으로 주제군집 수 t가 증가하고 하위군집 수 k가 감소할수록 코사인유사도가 감소한다는 사실을 통해 t 가 작고 k 가 큰 모수 설정이 필요함을 알 수 있다.

4-2. 두 번째 실험

세 번째는, 본 연구방법과 다른 방법들의 성능을 비교하였다. 성능 비교에 쓰인 모형은 *LDA* 과 *Doc2Vec*[5] 이다.

총 군집 수	*DOCST* (t = 16)	*LDA*	*D2V*
32	0.4446	0.3938	0.0827
48	0.5019	0.4954	0.0938
64	0.5377	0.4638	0.0857

<표 5> 3 가지 방법의 코사인 유사도

[표 5]를 통해 본 연구의 방법이 다른 방법에 비해 좋은 성능을 보여주고 있음을 확인 할 수 있다. *Doc2Vec* 의 경우, 충분하지 않은 데이터에 대해서 학습이 잘 되지 않는다는 것을 보여준다. 또한, *LDA* 와 비교해도 본 연구방법이 더 좋은 성능을 보인다. 이번 실험을 통해 짧은 문장에서 본 연구방법이 보다 좋은 결과를 나타낸다.

4-3. 세 번째 실험

세 번째 실험에서는 실제로 군집된 결과를 살펴보았다. [표 6]은 *LDA* 를 통해 64 개로 나누어진 군집 중 하나를 요약한 내용이다. 군집에 있는 문장이 모두 크게 다르지 않음을 알 수 있다.

군집	*LDA*
1	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO COONABARABRAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO COONABARABRAN (중략) QUIRINDI OR GUNNEDAH MAY BE

	ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH TIMES MAY VARY AT SHORT NOTICE PILOT RESPONSIBILITY TO CK AND MNT STS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN(중략) QUIRINDI OR GUNNEDAH TIMES MAY VARY AT SHORT NOTICE PILOT RESPONSIBILITY TO CK AND MNT STS

<표 6> *LDA*의 군집 내용 (총 군집 수: 64)

군집	*DOCST*
1	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
2	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO COONABARABRAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO COONABARABRAN (중략) QUIRINDI OR GUNNEDAH MAY BE ACT/DEACTIVATED AT SHORT NOTICE PILOT RESPONSIBILITY TO CK CURRENT STS WITH ATS
3	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH TIMES MAY VARY AT SHORT NOTICE PILOT RESPONSIBILITY TO CK AND MNT STS
	WILLIAMTOWN WEST AIRSPACE R559A ACT (중략) SYDNEY TO ANBAN (중략) QUIRINDI OR GUNNEDAH TIMES MAY VARY AT SHORT NOTICE PILOT RESPONSIBILITY TO CK AND MNT STS

<표 7> *DOCST*의 군집 결과 (총 군집 수: 64)

[표 7]은 [표 6]과 동일한 내용의 문서를 *DOCST*로 군집한 결과다. *DOCST*또한 64 개의 군집으로 나누었으며, 3 개의 군집에서 [표 6]과 동일한 문서가 발견되었다. 문장을 좀 더 자세히 살펴보면 3 개의 군집으로 묶인 문서들간의 차이점을 발견 할 수 있다.

[표 7]의 1 번 그룹은 ANBAN 이라는 단어와 BE ACT/DEACTIVATED 라는 단어가 포함되어 있다. 반면에 2 번 그룹은 1 번 그룹의 단어 ANBAN 대신에 COONABARABRAN 이라는 단어가 있음을 확인 할 수 있다. 3 번 그룹은 나머지 두 그룹과 다르게 BE ACT/DEACTIVATED 대신 MAY VARY 라는 단어가 있음을 확인 할 수 있다.

5 장. 결론

본 논문은 잡음이 포함된 단문에 대해서 가중치 *SDF*를 이용한 효과적 군집 방법인 *DOCST*를 제안한다. 단문 데이터를 일반적인 방법으로 군집할 경우, 자료의 희박함(Sparsity) 문제로 분석에 어려움이 발생한다. 본 연구는 이러한 어려움을 극복하고자 *LDA*를 통해 희박함 문제를 일정부분 해소하였으며, *Word2Vec*가 잘 학습 될 수 있도록 하였다. 또한, *SDF*를 통해 각 단어 벡터의 위치를 조정하여 축약어가 혼재된 단문도 잘 군집 될 수 있도록 하였다.

본 연구는 연구자가 결정해야 하는 모수가 매우 많다. 디리클레 분포의 모수 α, β 와 주제군집 수 t, *Kmeans*의 하위군집 수 k 등 많은 모수를 초기에 설정해야 하는 어려움이 있다. 또한, *Kmeans*외에 다른 군집방법을 사용할 경우에 대한 추가 연구가 필요하다.

참고문헌

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan *“Latent Dirichlet Allocation”* Journal of Machine Learning Research 3 (2003) 993-1022

[2] Yoshua Bengio, R´ejean Ducharme, Pascal Vincent, and Christian Janvin. *“A neural probabilistic language model”* Journal of Machine Learning Research 3 (2003) 1137–1155

[3] Burnham, K. P.; Anderson, D. R. *"Multimodel inference: understanding AIC and BIC in Model Selection"* Sociological Methods & Research (2004) 33: 261–304

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, *“Distributed Representations of Words and Phrases and their Compositionality”* Advances in neural information processing systems (2013)

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *“Efficient Estimation of Word Representations in Vector Space”* International Conference on Learning Representations (2013)

[6] Quoc Le and Tomas Mikolov. *“Distributed representations of sentences and documents”* International Conference on Machine Learning (2014)

[7] Cedric De Boom, Steven Van Canneyt, Steven Bohez, Thomas Demeester, Bart Dhoedt *“Learning Semantic Similarity for Very Short Texts”* (2015)

[8] 이진원. *“LDA 모형에서 Hyper parameters 의 영향력에 대한 연구”* (2017)

[9] http://naver.me/FYh17iMM