소셜 미디어 시계열 예측을 이용한 고객 니즈의 부상성 탐지

: 워드 임베딩, 네트워크 분석, LSTM 기반의 부상 키워드 탐지 방법

김명섭*, 박영재, 이승주, 이권능, 최재은

RESEARCH BACKGROUND AND PURPOSE

1. 기술 기회 포착

- 새로운 기회를 포착하는 것은 기업의 존속에 있어 가장 중요한 일 중 하나
- 기업은 고객에게 새롭거나 향상된 제품을 제공하기 위해 고객의 목소리(VOC)에 주의를 기울여야 함

2. 소셜 미디어 분석

- 소셜 미디어는 제품에 대한 공개적인 의견을 교환하기 위한 매체
- 제품 수명 주기의 단축에 따라 소셜 미디어를 통한 역동적인 고객 니즈의 분석이 중요

3. 고객 니즈 조기 탐지

- 특히 향후 부상할 고객의 요구 사항(니즈)을 조기에 탐지하고 예측하는 것이 중요
- 이를 통해 기업은 경쟁 기업이 쉽게 모방할 수 없는 고객과의 관계를 구축하여 기업의 경쟁력 강화로 이어짐

-> 부상할 고객 니즈를 조기에 탐지하여 제품 개발에 반영할 수 있도록 함

PROPOSED METHODOLOGY

STEP 1 데이터 수집 & 전처리

- 데이터 수집

Google Big Query를 이용해 Reddit 사이트에서 40개월 동안의 게시글(post)와 댓글(review)를 수집

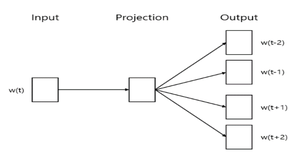
- 데이터 전처리

	doc_index	keyword	time
▶	87	understand	2015/12/31
	87	use	2015/12/31
	87	phillips	2015/12/31
	87	interface	2015/12/31
	87	is	2015/12/31
	87	just	2015/12/31
	87	lights	2015/12/31
	87	devices	2015/12/31
	87	directly	2015/12/31
	87	do	2015/12/31
	87	e	2015/12/31
	87	open	2015/12/31

게시글과 댓글의 텍스트를 띄어쓰기 단위로 분할 및 등록 월별로 분류

STEP 2 가중치 행렬 생성

- Word2vec 임베딩 및 가중치 적용



	ability	ac	access	...	zones
ability	1	0.87	0.73	...	0.12
ac	0.25	0.78	0.49	...	0.54
access	0.11	0.46	0.28	...	0.89
...	...	...	...	...	...
zones	0.27	0.39	0.97	...	0.08

Weight Matrix

Word2vec으로 생성된 임베딩 된 차원에서 키워드 간의 거리에 가중치를 적용하여 가중치 행렬 생성

$\omega_{ij} = \exp\left(-\frac{d(x_i, x_j)^2}{\sigma}\right) where\ node\ i\ and\ node\ j\ adjacent\ \ (eq.1)$

$d(x_i, x_j)$: $distance\ between\ node\ i\ and\ node\ j$

STEP 3 키워드 중요도 산출

- 가중치와 동시 출현을 고려한 closeness centrality 계산

	ability	ac	access	...	zones
ability	1	0.87	0.73	...	0.12
ac	0.25	0.78	0.49	...	0.54
access	0.11	0.46	0.28	...	0.89
...	...	...	...	...	...
zones	0.27	0.39	0.97	...	0.08

Weight Matrix

X

	ability	ac	access	...	zones
ability	1	2	1	...	1
ac	8	1	2	...	8
access	9	2	1	...	5
...	...	...	...	...	...

Co-occurrence Matrix

가중치 행렬에 동시 출현 행렬을 요소 별로 곱하여 단어 네트워크를 생성 후 계산된 closeness centrality를 키워드의 중요도로 간주

- $closeness\ centrality(x) = \frac{N}{\sum_y d(y,x)}\ (eq.2)$

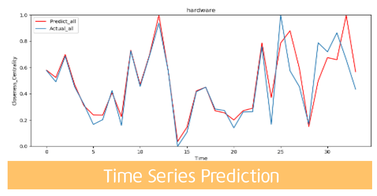
$d(y,x)$: $shortest\ distance\ between\ node\ i\ and\ node\ j$

STEP 4 키워드 중요도 예측

- 중요도의 시계열 예측

keyword	2015_12	2016_1	2016_2	2016_3	2016_4
the	0.94785	0.94747	0.94544	0.94436	0.94262
i	0.90915	0.92061	0.90892	0.92035	0.90968
to	0.92611	0.93589	0.92749	0.92666	0.92595
a	0.92023	0.92486	0.92643	0.92783	0.92299
and	0.91309	0.91515	0.91328	0.906	0.91186
it	0.87854	0.8848	0.88595	0.88168	0.88831
you	0.80757	0.82749	0.81186	0.8177	0.81997
that	0.85093	0.85912	0.85421	0.86048	0.86016
is	0.85634	0.86346	0.85757	0.85986	0.86555
of	0.86851	0.8672	0.85773	0.86373	0.86398
for	0.84844	0.84282	0.84485	0.85305	0.84912

Keyword Importance over Time

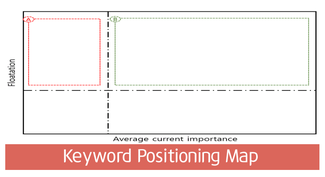


Time Series Prediction

시계열적 키워드 중요도를 input으로 LSTM을 적용하여 키워드의 중요도를 예측

STEP 5 부상 키워드 탐지

- 미래에 파급을 일으킬 키워드 탐지



Keyword Positioning Map

키워드의 중요도와 부상성을 이용하여 keyword positioning map을 구성하고 미래에 파급을 일으킬 고객의 니즈를 탐지

- $floatation(keyword) = \frac{avg(current\ importance)}{avg(predicted\ importance)}\ (eq.3)$

CASE STUDY

데이터 수집 및 전처리

- reddit에서 2015년 12월 부터 2019년 3월 까지 40개월간 sub reddit이 Home automation에 속하는 post 31,430개와 review 296,580개 수집

가중치 행렬 생성

- skip-gram방식 사용, 100차원 공간에 임베딩, 윈도우 크기 5로 설정, 신경망 반복 횟수(epoch) 1000번으로 설정,
- σ값은 키워드 간의 거리의 총 분산 사용

keyword	MSE
home	0.0178
switch	0.15143
hub	0.07157
control	0.04481
switches	0.0922
...	...
deployed	0.03786
average	0.07437

LSTM

Keyword	MSE
home	0.20215
switch	0.53033
hub	0.23493
control	0.19974
switches	0.15046
...	...
deployed	0.15671
average	0.1961

ARIMA

Keyword	MSE
home	0.25267
switch	0.20366
hub	0.1855
control	0.21102
switches	0.19074
...	...
deployed	0.12078
average	0.14049

Prophet

키워드 중요도 산출

- 2015년 12월부터 2019년 3월까지 1개월 단위로 모든 키워드에 대해 중요도를 계산
- the, I, things등 기술적으로 유의하지 않다고 판단되는 단어를 정성적으로 제거

키워드 중요도 예측

- 1개월 단위로 총 40개월의 데이터를 이용하여 향후 12개월의 키워드 중요도를 예측
- LSTM, ARIMA, Prophet의 키워드 평균 MSE를 비교하여 예측 모델을 채택

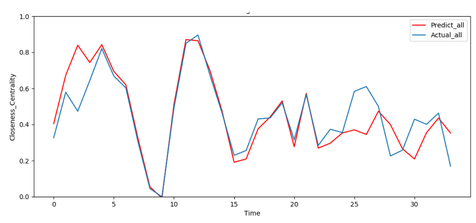
#LSTM 하이퍼 파라메터

time step = 6, batch size = 1
train : test = 75 : 25, layer 4층
hidden node = 16,
optimizer = adam
epoch = 1800, loss = MSE

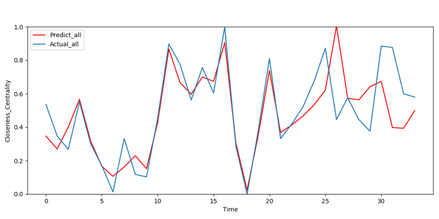
학습 결과

Predict
Actual

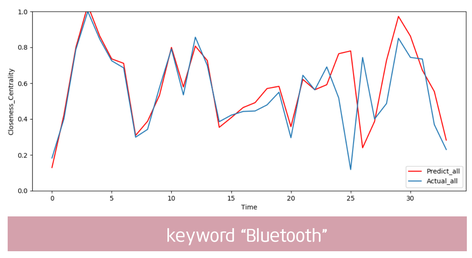
세로축: 중요도(Closeness centrality)
가로축: 시간(개월)



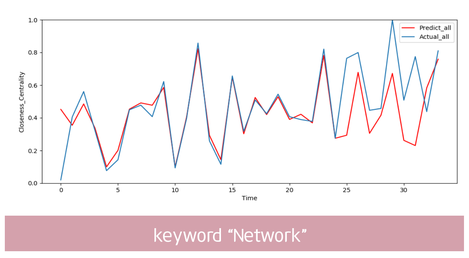
keyword "Integration"



keyword "Display"



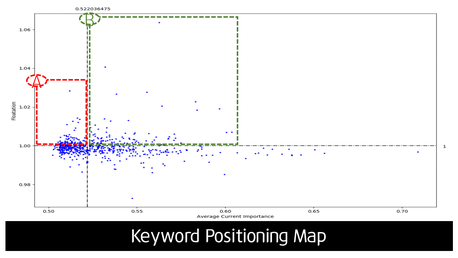
keyword "Bluetooth"



keyword "Network"

RESULT

부상 키워드 탐지



Keyword Positioning Map

- **X축** : 1개월 단위로 수집되어 계산된 총 40개월의 키워드 중요도 평균(**현재 중요도**)
- **Y축** : 예측된 12개월의 키워드 중요도 평균을 미래 중요도로 간주, 미래 중요도를 현재 중요도로 나눈 값인 **부상성**(Floatation)
- 각 키워드의 **현재 중요도**와 **부상성**을 이용하여 **키워드 포지셔닝 맵**을 구성, x축의 경우 경계 값을 중앙값(median), y축의 경우 경계 값을 1로 선정

keyword	floatation	keyword	floatation	keyword	floatation	keyword	floatation
customization	1.02831502	error	1.00418259	programmed	1.00250865	linking	1.00173135
lan	1.01433034	bug	1.00416566	communicates	1.00242058	notified	1.00167877
circuits	1.01365671	recognition	1.00407463	microphone	1.00240875	latency	1.0016484
maintenance	1.01263222	coverage	1.00406242	reconnect	1.0023799	smarthings	1.00158705
arlo	1.01214871	repeater	1.00404148	sw	1.00235627	cabinet	1.00158658
protection	1.01099093	traffic	1.00381006	homebridge	1.00222735	netflix	1.0015266
bedrooms	1.00909272	projector	1.00379006	siren	1.00220738	availability	1.00152305
bandwidth	1.00797692	adapters	1.00368134	micro	1.00214595	smartapp	1.00150937
warranty	1.00576672	cars	1.00356644	consumers	1.0021202	heaters	1.00150703
cabinets	1.00537371	charging	1.00345734	minimote	1.00193154	spotify	1.00150392
chromecasts	1.00535521	decora	1.003232	speeds	1.00192895	detecting	1.00145649
charger	1.00487636	versions	1.00314513	stereo	1.00191302	robust	1.00141589
isp	1.00484173	damage	1.00314076	upload	1.00188367	construction	1.00141114
bell	1.00460183	centralized	1.00291827	listening	1.00187256	engine	1.00139358
aux	1.00438445	package	1.00265246	itunes	1.00177302	rental	1.00130759

Keywords of Area A

- A 구역의 키워드를 부상성이 큰 순서로 나열
- "customization", "lan", "circuits", "maintenance", "protection" 등
- 해당 키워드들은 현재는 낮은 중요도를 가지나 향후 파급을 일으킬 것으로 예상되는 높은 부상성을 가짐

"I recently got a Phillips Hue Motion Sensor, and while I'm impressed by how well it works, I'm rather disappointed by the level of **customization**/automation. "

"I've been using HS3 since 2014 and love it for it's stability and **customization**."

Customer Needs of Keyword "Customization"

- "최근에 Phillips Hue Motion Sensor를 구매했는데 잘 작동되어 좋은 인상을 받았지만 사용자 지정(customization)/자동화 수준에 다소 실망했다."
- " 2014년부터 HS3을 사용해 왔으며 안정성과 사용자 지정(customization)이 마음에 든다."

keyword	floatation	keyword	floatation	keyword	floatation	keyword	floatation
camera	1.063719354	chromecast	1.008557094	option	1.004616325	settings	1.003310842
batteries	1.040615991	mac	1.007268724	wemo	1.004591425	outlet	1.003296849
replace	1.027635834	bathroom	1.007116059	wired	1.004580374	radio	1.00315914
remotely	1.026632446	setup	1.0070579	computer	1.00452459	cheaper	1.003118392
assistant	1.02268818	room	1.006947609	car	1.004281208	toggle	1.003051328
cameras	1.020446257	router	1.00662251	speed	1.004141929	wiring	1.002969196
card	1.02036392	ecobee	1.006532205	cable	1.003978392	tv	1.002828434
wifi	1.019047867	wife	1.006386376	echo	1.003926455	siri	1.002728739
alexa	1.018444186	reliable	1.006059207	tell	1.003797766	replacing	1.002711603
automations	1.013887802	connecting	1.006020214	sound	1.003791696	manual	1.002702266
area	1.01311721	installed	1.006013098	address	1.003693641	ios	1.002643448
moving	1.012776049	cables	1.005618274	action	1.003635734	wires	1.002550054
gateway	1.010658058	page	1.005517622	module	1.003605629	receiver	1.002545614
storage	1.009498815	ecosystem	1.004962104	wireless	1.003516689	firmware	1.00250411
mini	1.009333461	plugs	1.00464113	ceiling	1.003428765	bridge	1.002489427

Keywords of Area B

- B 구역의 키워드를 부상성이 큰 순서로 나열
- "camera", "batteries", "replace", "remotely", "wifi"등
- 해당 키워드들은 현재 높은 중요도를 가지며 향후 더욱 중요해질 것으로 예상되는 키워드

"Might need **camera**s after all but I wonder how much a camera costs that can see well enough in the dark."

"I really love the Floodlight **camera**. I'd probably get one in a heartbeat if it had an available local video stream."

Customer Needs of keyword "Camera"

- "결국 카메라(camera)가 필요할지 모르지만 어둠 속에서 충분히 볼 수 있는 카메라 비용이 얼마인지 궁금하다."
- "플러드 라이트 카메라(camera)가 정말 마음에 든다. 로컬 비디오 스트림을 사용할 수 있다면 경우 기꺼이 구매할 것이다."

-> 해당 키워드들은 고객 니즈를 조기에 파악하고 예측하여 경쟁사 대비 우위를 점하기 위해 주시가 필요한 키워드

CONCLUDING REMARKS

연구 의의

본 연구에서는 소셜 미디어 예측을 이용해 정량적이고 시스템적인 방법으로 향후 부상할 고객의 니즈를 파악하는 방법을 제안

기대 효과

전문가의 판단에 비해 적은 시간, 낮은 비용으로 기업의 경쟁력 강화를 위한 정량적인 의사결정을 지원하는데 기여할 것으로 기대