빅데이터를 활용한 직업관련 분석 및 시각화

최동철[O], 최낙진[*], 김민석[*], 박준욱[*], 이준동[*]
[*]강릉원주대학교 멀티미디어공학과,
[O]강릉원주대학교 멀티미디어공학과
e-mail: dong095362@naver.com[O], choi@ex.co.kr[*], ggspsp96@naver.com[*],
jaeminet@gmail.com[*], jlee@gwnu.ac.kr[*]

Job-related analysis and visualization using big data distributed processing system

Dong-Cheol Choi[O], Nakjin Choi[*], Min-Seok Kim[*], Jun-wook Park[*], Jun-Dong Lee[*]
[*]Dept. of Multimedia Engineering, GangNeungWonju University,
[O]Dept. of Multimedia Engineering, GangNeungWonju University

● 요 약 ●

본 논문에서는 코로나바이러스감염증19 사태가 국내 취업시장에 어떠한 영향을 미쳤는지에 대해 알아보기 위하여 빅데이터를 활용한 직업 관련 분석 및 시각화를 수행하였다. 빅데이터를 위한 기본 자료는 통계청 자료와 워크넷 Open API를 활용하였으며, 빅데이터 처리 과정을 거쳐 결과값을 예측을 시도하였다. 2020년도 워크넷 Open API를 통해 고용수와 통계청 자료를 통해 비교 분석 및 시각화를 실시하였고, 08년~20년 취업자수를 통해 시계열 분석 및 예측을 진행해 앞으로의 횡보를 예상해보았다. 분석한 결과 19년, 20년도를 비교 분석했을 때에는 크게 차이가 나지 않았다. 추가적으로 시계열 분석기법을 활용해 보았을 때 매년 고용수는 전체적으로 증가하고 4월에는 감소, 7월에는 증가하는 추세가 나왔다. 코로나바이러스감염증19 사태로 인해 공공기관과 언택트 시대에 따른 화상회의나 재택근무로 인해 운수•통신 취업률은 상승한다는 결과값이 도출되었고, 자영업이나 서비스 직업 등은 다른 직종에 비해 큰 감소를 보여줬으나 국가 경제 활성화에 따른 고용수는 점차 증가할 것이라 예측된다.

키워드: 빅데이터(Big data), 반정형 데이터(semi-structured data), 정형 데이터(Structured data), Open API(Open Applic ation Programming interface), 데이터 분석(Data analysis), 서버(Server), 크롤링(Crawling)

I. Introduction

2020년1월20일 국내 첫 확진자가 발생한 후 코로나바이러스감염증19가 확산되면서 국민의 일상생활부터 국가와 경제·사회에 영향을 미치고 있다. 본 논문에서는 코로나바이러스감염증19 관련해서 취업시장에 어떠한 영향을 미치는 지에 대해 분석 및 시각화, 시계열 분석, 예측을 진행하게 되었다.

II. Preliminaries

빅데이터란 기존의 자료처리를 위한 소프트웨어로 적절하게 처리하기 힘든 방대한 양의 데이터로 정의하고 있다. 빅데이터 활용에는 공공서비스 분야와 일반서비스 분야로 나누어지는데 본 논문에서는 일반 서비스 분야로 빅데이터를 사용한 주제를 분석 및 예측을 하기 위해서 Fig. 1. 과 같은 과정을 통해서 진행하게 되었다. 빅데이터는 총 5가지 수집-저장-처리-분석-시각화의 과정을 거친다[1][2][3].

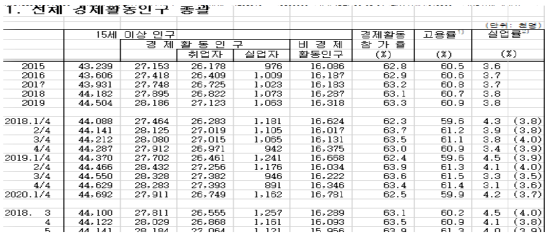
Fig. 1. Big Data Processing Process

III. The Proposed Scheme

1. 데이터 수집

- 통계청 고용 동향 Dataset

통계청 홈페이지에서 대한민국 고용동향에 대한 자료(Fig. 2.), 워크넷 OpenAPI(Fig. 3.)를 통해서 20년 3~5월 4만 여개의 고용모집 데이터를 크롤링하여 분석에 필요한 자료 수집을 실시하였다.



	15세 이상 인구	경제활동인구			비경제활동인구	경제활동참가률 (%)	고용률 (%)	실업률 (%)	
			취업자	실업자					
2015	43.239	27.153	26.178	976	16.086	62.8	60.5	3.6	
2016	43.606	27.418	26.409	1.009	16.187	62.9	60.8	3.7	
2017	43.931	27.748	26.725	1.023	16.183	63.2	60.8	3.7	
2018	44.182	27.995	26.922	1.073	16.287	63.1	60.7	3.8	
2019	44.504	28.186	27.123	1.063	16.318	63.3	60.9	3.8	
2018.1/4	44.088	27.464	26.283	1.181	16.624	62.3	59.6	4.3	(3.8)
2/4	44.141	28.125	27.019	1.105	16.017	63.7	61.2	3.9	(3.8)
3/4	44.212	28.080	27.015	1.065	16.131	63.5	61.1	3.8	(4.0)
4/4	44.287	27.912	26.971	942	16.375	63.0	60.9	3.4	(3.9)
2019.1/4	44.370	27.702	26.461	1.241	16.668	62.4	59.6	4.5	(3.9)
2/4	44.466	28.432	27.256	1.176	16.034	63.9	61.3	4.1	(4.0)
3/4	44.550	28.328	27.382	946	16.222	63.6	61.5	3.3	(3.5)
4/4	44.629	28.283	27.393	891	16.346	63.4	61.4	3.1	(3.6)
2020.1/4	44.692	27.911	26.749	1.162	16.781	62.5	59.9	4.2	(3.7)
2018. 3	44.100	27.811	26.555	1.257	16.289	63.1	60.2	4.5	(4.0)
4	44.122	28.029	26.868	1.161	16.093	63.5	60.9	4.1	(3.8)

Fig. 2. Uses Dataset-1

구인인증번호	회사명	채용제목	임금형태	급여	최소임금액	최대임금액	근무지역	등록일자	마감일자	직종코드	직업분류
K120032005210004	(주)로나바	임원채용	연봉	4500만원	45000000	85000000	서울 강남	20-05-21	채용시까지	011200	기업 고위임원
K131212005150009	(주)드림엔(경력)화공	철금	500만원	5000000	0		울산 남구	20-05-15	채용시까지	011200	기업 고위임원
K141222004030003	주식회사	대신브레이	연봉	3000만원	30000000	40000000	경북 김천	20-04-03	채용시까지	011200	기업 고위임원
KC15442005210001	주식회사	'관리팀 직	철금	180만원 ~	1800000	1800000	전남 순천	20-05-21	채용시까지	012200	경영지원 관리자
K172222005190014	한내시장	보행한내시	철금	250만원 ~	2500000	3200000	충남 보령	20-05-19	20-05-31	012200	경영지원 관리자
K140132005190021	(주)휴먼피	본사 영업	연봉	4000만원	40000000	0	대구 동구	20-05-19	채용시까지	012200	경영지원 관리자
KJLA002005180001	재단법인	'2020년 영	철금	294만원	2947200	0	강원 영월	20-05-18	20-06-05	012200	경영지원 관리자
K120232005140021	(주)삼안시	종합건설	연봉	3000만원	30000000	45000000	서울 강동	20-05-14	채용시까지	012200	경영지원 관리자
K162212005130020	(주)위드피	재무회계	연봉	4000만원	40000000	0	전남 순천	20-05-13	20-05-31	012200	경영지원 관리자
K151822005130027	(주)에스아	신재생에너	연봉	3000만원	30000000	30000000	경기 파주	20-05-13	채용시까지	012200	경영지원 관리자
K170052005120044	재단법인	대전세종인	연봉	6178만원	61788000	0	대전 중구	20-05-12	20-05-29	012200	경영지원 관리자
K120032005120039	에스앤플	경영지원차	연봉	4000만원	40000000	50000000	서울 강남	20-05-12	20-06-01	012200	경영지원 관리자

Fig. 3. Uses Dataset-2

2. 데이터 분석

통계청 고용동향 과 워크넷 OpenAPI 데이터셋을 전처리 작업을 통해 데이터 처리를 먼저 진행 한 후, 다양한 분석 기법을 표현 및 예측을 할 수 있도록 파이썬을 활용하여 설계를 진행 하였다.

```
import pandas as pd
import plotly.graph_objs as go
import plotly.offline as py
from fbprophet import Prophet
from fbprophet.plot import plot_plotly, add_changepoints_to_plot
import numpy as np

df = pd.read_csv('C:/MUser/MMASTERMM/pythonMM/training.csv',encoding='cp949', thousands = ',')
df = df.drop(df.index[[-40]])
df['date'] = pd.to_datetime(df['date'])

data = df[['date', 'work']].reset_index(drop=True)
data = data.rename(columns={'date': 'ds', 'work': 'y'})

fig = go.Figure()
fig.add_trace(
    go.Scatter(
        x=data.ds,
        y=data.y,
        name="Unfilled in Korea"
    )
)
fig
df_prophet = data.rename(columns={
    'date': 'ds',
    'work': 'y'
})
forecast = m.predict(future)
forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail(7)
m = Prophet(
    changepoint_prior_scale=0.2,
    changepoint_range=0.95,
    yearly_seasonality=True,
    weekly_seasonality=False,
    daily_seasonality=True,
    seasonality_mode='additive'
)
m.fit(df_prophet)
future = m.make_future_dataframe(periods=365)
forecast = m.predict(future)
fig = plot_plotly(m, forecast)
py.iplot(fig)
fig = m.plot(forecast)
a = add_changepoints_to_plot(fig.gca(), m, forecast)
fig2 = model.plot_components(forecast)
```

Fig. 4. Analysis and prediction code

3. 데이터시각화

코로나 사태가 일어난 이후 시간과 취업률에 따른 변화를 조금 더 다양하게 보여줄 수 있도록 수집된 데이터셋을 활용하였다. 먼저 데이터셋을 분류하여 분석 및 예측을 실행하였다. 먼저 가장 많이 언급된 데이터를 활용하여 시각적으로 차이를 비교하기 편한 fig. 5. 와 같은 그래프로 표현을 하였고, 앞으로 예측할 데이터 또한 시계열 분석을 통해 Fig. 6.와 같은 분석 그래프를 나타내었다.

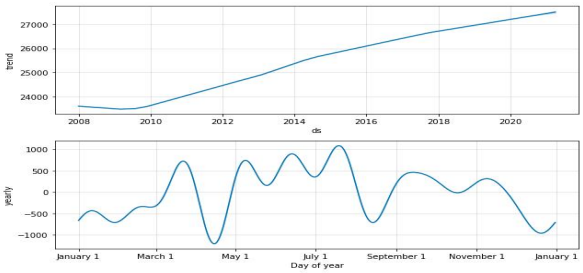


Fig. 5. Visualization graph

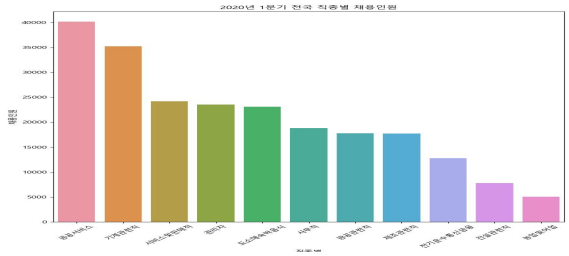


Fig. 6. Time series analysis and forecast graph

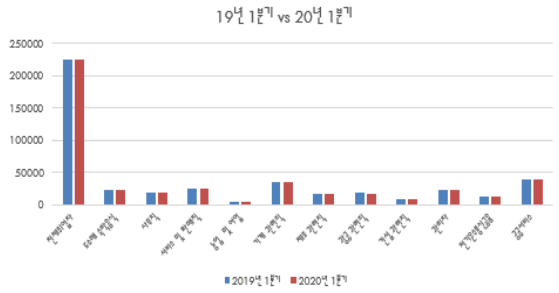
IV. Conclusions

본 논문은 빅데이터 분석과정을 통해 Fig. 7. Fig. 8.와 같이 분석 및 예측 결과 값이 나왔다.

19년도 데이터와 큰 폭은 없으며, 평균적으로 매년 4월은 고용수가 감소, 7월은 증가하는 것을 볼 수 있다.

코로나바이러스감염증19 사태로 인해 공공기관(병원 등)과 언택트 시대에 따른 화상회의나 재택근무로 인해 운수•통신 취업률은 상승한다는 결과 값이 도출되었다.

자영업이나 서비스 직업 등은 다른 직종에 비해 큰 감소를 보여줬으나 국가 경제 활성화에 따른 고용수는 점차 증가할 것이라 예측된다 [10].



1년 1분기 vs 20년 1분기												
(인원)	전매위업자	도소매 숙박음식	사무직	서비스 및 판매직	농업 및 어업	기계 관련직	재조 관련직	광공 관련직	건설 관련직	관리자	전기운수통신금융	공공서비스
2019년 1분기	225679	23776	18901	24334	4653	34805	17717	17777	7903	23728	12606	39479
2020년 1분기	226063	23064	18824	24184	5053	35204	17692	17751	7819	23543	12758	40171
(인원 차분치)	전매위업자	도소매 숙박음식	사무직	서비스 및 판매직	농업 및 어업	기계 관련직	재조 관련직	광공 관련직	건설 관련직	관리자	전기운수통신금융	공공서비스
2019년 1분기	225679	10.535	8.375	10.783	2.062	15.422	7.851	7.877	3.502	10.514	5.586	17.493
2020년 1분기	226063	10.202	8.327	10.698	2.235	15.573	7.826	7.852	3.459	10.414	5.644	17.770
변화 율(인원)	-384	0.333	0.048	0.085	-0.173	-0.150	0.024	0.025	0.043	0.100	-0.058	-0.276

Fig. 7. Visualization graph results

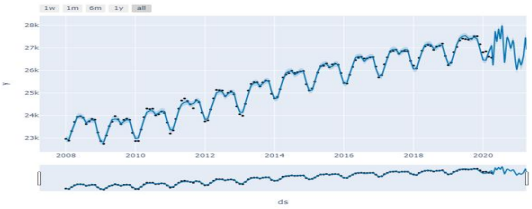


Fig. 8. Time series analysis results

ACKNOWLEDGMENT

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단 - 현장맞춤형 이공계 인재양성 지원사업의 지원을 받아 수행된 연구임(NRF-2017H1D8A1029391).

REFERENCES

[1] Hanol, “Introduction to Big Data Analysis, R and Big Data Platform, Hadoop” pp.482-485, February 2020.
[2] Seong-Hoon Lee, “Big Data Usage Status” pp.51-54, September 2019.
[3] Multicampus, "Big Data Distributed Parallel Processing System and Introduction to Analysis" pp. 9-30, February 2020.
[4] http://kostat.go.kr/portal/korea/index.actionStatistical (Office Employment Trends OpenData)
[5] https://openapi.work.go.kr/opiMain.do (Worknet OpenAPI)
[6] https://www.fun-coding.org/crawl_basic2.html (Web Crawling)
[7] https://wikidocs.net/16582 (Pretreatment analysis)
[8] https://zzsza.github.io/development/2018/08/24/data-visualization-in-python/ (Python Visualization)
[9] https://zzsza.github.io/data/2019/02/06/prophet/ (Facebook Time series analysis code)
[10] Bae Young Im, Shin Hye-ri, “Corona 19 speeds up the un-contact society” pp.1-26, May 2020.