ISSN:1226-7244 (Print)
ISSN:2288-243X (Online)
논문번호17-01-03
j.inst.Korean.electr.electron.eng.Vol.21,No.1,13~23,March 2017
http://dx.doi.org/10.7471/ikeee.2017.21.1.13

포탈의 검색 트렌드를 활용한 인천공항 출국자 수 예측 연구

Search Trend's Effects On Forecasting the Number of Outbound Passengers of the Incheon Airport

신의섭*, 양동헌**, 손세창**, 허문행***, 백석철***★

Euiseob Shin*, Dong-Heon Yang**, Sei Chang Sohn**, Moonhaeng Huh***, Seokchul Baek***★

Abstract

Short-term prediction of the number of passengers at the airport is very essential for the efficient and stable operation of the airport. Here, to forecast the immigration of Incheon International Airport, we perform the predictive modeling of Korean and Chinese outbound travelers comprising most of immigration. We conduct the Granger Causality test between the number of outbound travelers and related search trend data to confirm the correlation. It is found that the forecasting with both "outbound travelers" and "search term trends" data outperforms the one only with "outbound travelers" data. This is because search activities are done before doing something and this study confirms that search trend data inherently possess the potential for prediction.

요 약

공항의 안정적인 운영을 위하여 승객의 단기예측은 매우 중요하다. 본 논문에서는 인천공항의 출입국자 예측을 위하여 출입국자의 대부분을 차지하는 한국인과 중국인의 출국자의 예측 모델링을 수행하였다. 예측 모델링 정확도 향상을 위해 네이버와 바이두 검색 트렌드 데이터를 활용하였다. 출국자 수들과 관련 검색 트렌드 데이터 간 Granger Causality 테스트를 수행하여 상관관계가 있음을 확인하였다. "출국자 수" 단독으로 예측하는 것보다 "출국자 수"와 "검색어 트렌드" 자료를 합하여 예측하는 것이 정확도가 향상됨을 알 수 있었다. 이는 검색이 어떤 일을 수행하기 전에 하는 행위이기 때문이고, 검색 트렌드 데이터 내에 태생적으로 예측 기재가 존재함을 본 연구를 통하여 확인할 수 있었다.

Key words : Big Data, Search Trend, Forecast Modelling, Linear Regression, Neural Network.

* Dept. of Electronic Eng., Soonchunhyang Univ.
** R&D Center, Incheon International Airport Corp.
*** Dept. of Digital Media, Anyang Univ.
Corresponding Author
e-mail: scbaek86@gmail.com, tel: 02-961-7611
Manuscript received Mar. 13, 2017;
revised Mar. 27, 2017 ; accepted Mar. 29, 2017

Ⅰ. 서론

항공운송 산업은 막대한 선행투자가 이루어져야 하기 때문에 항공 수요를 예측하는 것은 적절한 공항운영을 위해 필수적이다. 정부기관 및 항공운송사업자는 정책수립 및 사업수행에 앞서 신뢰성 있는 수요예측을 토대로 정책 및 경영계획 등을 수립해야 하며 만약 과학적 근거가 부족하고 잘못된 수요예측 방법을 적용할 경우 해당 항공정책

및 사업들은 과대 또는 과소투자를 발생시킬 가능성이 있다[1]. 공항의 설립, 증축 등 공항의 최대처리 용량과 관계된 예측은 장기 예측이 필요하고 공항의 스케줄링, 유지보수, 보안, 교통정책 등 안정적인 운용을 위해서는 단기 예측이 필요하다[2].

최근의 항공시장은 항공자유화, 저비용항공사의 성장, 전략적 제휴의 확대, 주변국 공항 개발 등으로 국가 간 경쟁이 심화되면서 대내외 항공운송 환경과 국가 정책의 변화를 적기에 반영하기 위해서는 지속적이고 신뢰성 있는 수요 예측이 요구되고 있다. 그러나 항공정책 수립에 기준이 되는 “항공정책 기본계획” 및 “공항개발 중장기 종합계획” 등의 항공 수요는 수송실적 위주의 단순 통계 모델을 반영하고 있고, 5년 주기로 발표되어 급변하는 환경변화에 대응이 어렵다. 또한 항공수요는 갑작스런 경기변동, 유가 및 환율 변동, 국내외 다양한 항공정책변화 등의 외부적 요인들로 인한 수요변화가 심화되고 있으므로 항공운송 특성과 환경변화를 감안한 지속적인 모니터링을 통해 국가차원의 체계적인 항공수요 관리가 필요하다[1]. 특히 안정적인 서비스를 위한 단기 예측은 수많은 변수가 존재하기 때문에 안정적인 수요예측 모델을 갖지 못하고 항공사에서 제공하는 예약 자료와 계절적 요소를 반영한 단순 통계 모델을 기반으로 약 1 ~ 2주 정도의 초단기 예측에 머물고 있다.

본 연구는 약 2년간 인천국제공항의 한국인 출국자와 중국인 출국자 자료를 바탕으로 단기간의 출국자를 예측하였고, 포탈에서 제공하는 검색 트렌드 자료를 추가로 활용하여 예측의 정확도를 높였다. 시계열 자료 간의 상관관계를 정량적으로 파악할 수 있는 통계기법인 “Granger Causality Test”를 수행하여 검색 트렌드와 한국인 출국자 사이의 상관관계가 있음을 확인하고 예측 모델링을 수행하여 검색 트렌드가 출국자 예측의 정확도를 높이는 예측기재로 활용됨을 검증하였다. 선형회귀분석 모델링과 신경망 예측 모델링을 활용하여 “출국자 수” 자료만을 사용한 예측, “출국자 수” + “검색 트렌드” 자료를 혼합 사용한 예측을 수행하였다. 예측 기법에 따라 “출국자” 혹은 “출국자 수” + “검색 트렌드”의 예측률의 우위가 달라짐을 확인할 수 있었다. 6개월 정도의 단기 예측의 경우, 두 가지 예측 방법 모두 “출국자 수” + “검색 트렌드”일 경우 예측 정확도가 높았다. 이는 포탈 검색 트렌드가 출국자를 예측할 수 있는 기재가 됨을 증명하는 것으로서 향후 다양한 예측 모델링에 포탈 검색 트렌드를 활용할 수 있다는 것을 의미한다.

본 논문의 구성은 다음과 같다. 2장에서 항공수요 예측에 관련한 연구를 소개한다. 3장에서 출국자를 예측하기 위한 자료 구성과 예측 기법을 소개하고 예측 기법을 활용하여 추출된 결과를 소개한다. 4장에서 연구결과에 대한 고찰을 하고 마지막으로 본 예측 기법을 활용한 장단점과 향후 연구 방향 등에 대하여 기술한다.

II. 관련연구

항공수요 예측 기법은 여러 가지가 있으며, 예측 용도에 따라 적용 방법은 달라진다. 정성적인 예측 방법은 시장조사(Market Survey), 델파이 기법(Delphi Method), 전문가 의견조사(Expert Method)를 사용하고 정량적인 예측 방법은 시계열분석 방법(Time Series Analysis)과 원인분석(Causal Analysis)으로 나뉜다. 또한 예측 기법은 예측하고자 하는 기간에 따라 적용하는 방법이 달라진다. 단기 예측은 1년 이내의 기간으로 현재 정책을 계획하거나, 실시중인 정책에 대한 평가 및 스케줄링과 같은 일상적인 업무와 연관되었을 때 필요하다. 중기 예측은 2 ~ 5년 정도의 기간으로 시장 계획 등에 적용을 위해 사용하며 장기 예측은 5년 이상의 기간으로 공항의 확장, 신축 등 대규모 예산이 수행될 것에 대비한 예측으로 사용한다[3].

시계열분석법은 항공여객 수요가 일정한 패턴을 형성하고 있다는 가정 하에 수행한다. 시계열 중에 급격한 변화(spike)가 있으면 예측이 힘들어지기 때문에 대부분 급격한 변화를 제거(smoothing)하고 예측을 수행한다. 일변량 시계열 모델 중 수요예측에 이용되는 분석 방법 중에서 가장 많이 쓰이는 모델은 Box & Jenkins의 계절형 자기회귀이동평균(ARIMA :Autoregressive Integrated Moving Average) 모델이다. Y. Kim 등은 ARIMA 모델을 사용하여 저가항공에 대한 수요 예측을 수행하였다[2]. Wolters는 시계열 자료를 exponential smoothing을 수행하고 ARIMA

모델을 사용하여 2008년 ~ 2020년 사이의 Lisbon 공항의 승객을 예측하는데 사용하였다[4]

다변량 시계열 분석은 일변량 시계열 분석에서 알 수 없는 시계열 변수들 사이에서 상호작용과 동적 관계를 설명하고 모델링하기 위해 사용한다. 다변량 시계열 모델은 VAR(Vector Auroregressive), VARMA(Vector Autoregressive Moving Average) 모델이 있다. J. Yoon 등은 계절형 다변량 시계열 모델을 이용하여 국제항공 여객 및 수요예측에 관한 연구를 수행하였다[5]. S. Baik 등은 제주-내륙 간 국내선 항공여객수요모델을 월별 시계열 총량자료를 이용하여 단순시계열 모델과 부분조정모델로 추정한 후 모델별 탄력성을 산출하였다[6].

원인분석법은 회귀분석법(Regression Analysis)와 중력분석법(Gravity Analysis)로 나뉜다. Duval 등은 각국의 GDP(Gross Domestic Product)와 뉴질랜드와 각국 간의 통화 교환 비율을 이용하여 뉴질랜드 승객을 예측하는데 회귀모델을 사용하였다[7]. 중력분석법은 뉴튼의 중력법칙을 이용하여 예측하는 기법으로 두 도시간의 인구를 중력으로 대신하여 예측한다. Grosche 등은 중력분석법을 이용하여 베를린과 유럽의 28개 도시간의 항공 승객 수를 예측하였다[8]

최근에는 대용량 자료 처리 기술의 발전에 따라 빅데이터 처리 기법을 활용한 예측분석도 많이 사용한다. S. Kim 등은 포탈의 검색엔진을 검색한 빈도를 활용하여 승객 수를 예측하였다[9]. S. Kim의 논문에서는 네이버의 검색 트렌드와 광고자료를 활용하여 자료를 수집하였고, "k-fold cross validation" 기법을 사용하여 분석하였다. Johan Bollen은 트위터 자료를 수집하여 사용자들의 기분(Mood)을 분석하고 이 기분과 예전의 주가를 합하여 주가의 부침을 예측하였다[10]. 즉 인터넷에서 수집한 사용자의 자료가 주가 예측 기재가 됨을 보였다.

III. 인천공항 출국자 예측

1. 예측 방법론

그림1은 본 논문에서 사용한 예측 순서도이다. 이 순서도는 Johan Bollen[10]에서 제시한 순서를 준용하였다. 그림1 (a), (b)와 같이 한국인 출국자 수를 예측하기 위해 2015. 1. 29 ~ 2016. 11. 29까지 인천공항을 통해 출국한 한국인과 중국인 출국자 수의 자료와 같은 기간 동안 한국의 네이버를 통해 "하나투어"를 검색한 빈도와 중국의 바이두(Baidu)를 통해 "서울"을 검색한 빈도를 활용하였다.

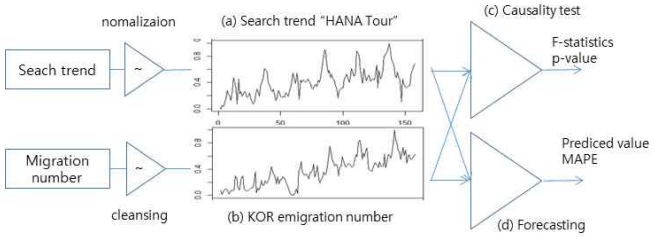

Fig. 1. Forecasting methodology(modified in [10])
그림 1. 예측 방법론([10]의 그림 수정)

그림. 1 (c) 는 (a)와 (b) 자료 간의 상관관계를 분석하는 과정이다. 그림1 (d)는 (a)와 (b) 자료를 활용하여 예측하는 과정이다. 본 논문에서는 (b) 자료만을 활용한 예측과 (a) + (b) 자료를 활용한 예측을 실시하였다.

2. 키워드 선택

인터넷 활용이 증가함에 따라, 많은 사람들이 어떤 일을 시작하기 전에 관련 일에 관해 검색을 하여 정보를 얻고 분석한 후 시작하는 경향이 뚜렷하다. 따라서 인터넷을 검색한 키워드에는 어떤 일에 대한 예측 기재가 포함되어 있을 수 밖에 없다. 본 논문에서는 인천공항을 통한 출국자 예측의 정확도를 높이기 위하여 출국자 자료와 포탈에서 검색한 키워드의 트렌드를 활용하였다. 한국에서는 네이버가 한국의 검색 시장의 80% 이상을 차지하고 있기 때문에 네이버의 검색 트렌드를 활용하였고[9][11] 중국에서는 바이두의 인덱스[12]를 활용하였다.

어떤 키워드가 출국자를 예측하는데 가장 효과적인지는 알려져 있지 않다. 본 논문에서는 여행을 위한 출국이 인천공항을 통한 출국자의 많은 부분을 차지한다는 것에 기인해 여행관련 키워드에 대하여 여러 가지 키워드를 입력하여 추출하였다. 한국인 출국자 예측에는 "하나투어" 키워드를 사용하였는데, 그림. 2와 같이 하나투어가 국내 여행사 중 주가 총액이 가장 크며 여행사 관련 키워드 중 가장 높은 순위를 차지하고 있었다.

그림2는 2015. 1. 29 ~ 2016. 11. 29까지 “하나투어”, “모두투어”, “롯데관광”에 대한 검색 트렌드에 대한 상대 비교이다.

인천공항의 중국인 출국자 예측을 위하여 “서울” 키워드를 사용하였다. “한국여행”, “한국비자”, “K-POP” 등의 키워드를 활용하였으나 만족할 만한 예측기재가 되지 못했다.

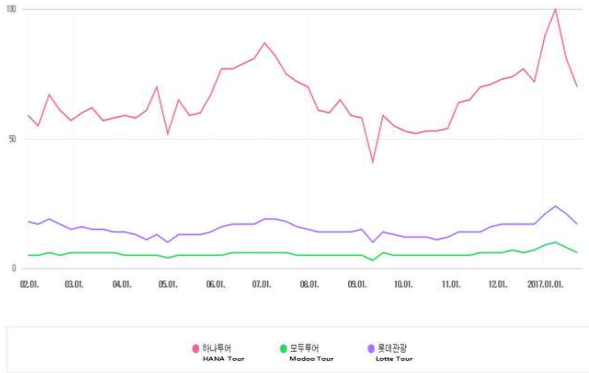

Fig. 2 Comparison of "Tour" relate Search trend
그림 2 여행사 관련 검색 트렌드 비교

3. 한국인 출국자 자료

그림. 3에 2015. 1. 29 ~ 2016. 11. 29 기간 동안의 한국인 출국자 자료를 보였다. 점차 증가하는 추세를 보이며 시계열 가운데 많은 스파이크(spike)가 존재한다. 이 스파이크는 설과 추석 등의 연휴에 발생한다. 연휴 전에는 큰 폭으로 출국자가 감소하고 연휴 시작과 함께 크게 증가하는 경향을 보인다. 이런 스파이크를 휴일효과(holiday effect)라고 한다. 이 스파이크는 예측 모델링에 상당히 어려움을 제공한다. 이런 휴일효과를 제거한 연구도 진행되었고, 예측 성능이 상당히 증가되었다고 한다[13]. 본 논문에서는 스파이크를 제거하지 않고 예측을 진행하였다. 그리고 확보된 자료의 전반 90%를 학습 자료로 활용하고 나머지 후반 10%를 테스트 자료로 활용하였다.

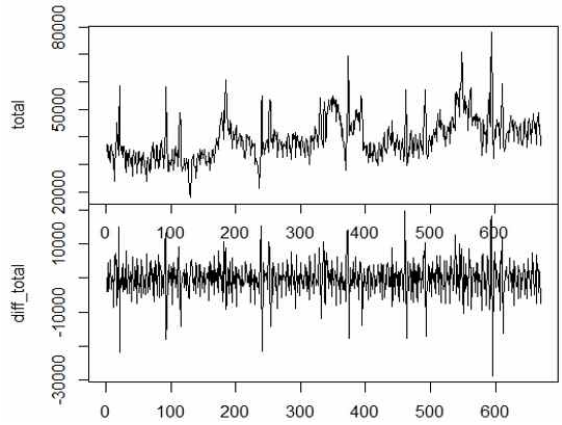

Fig. 3 (upper) Graph of Korean emigration number through Incheon Airport, (lower) difference of Korean emigraion number KOR(t) - KOR(t-1)
그림 3. (위) 인천공항을 통한 한국인 출국자 수
(아래) 출국자수 변화 그래프 KOR(t) - KOR(t-1)

4. 상관관계 분석

그림1의 (a)와 (b)에 제시한 서로 다른 시계열 자료 사이에 시차를 둔 상관관계가 존재하는지 조사할 필요가 있다. 시계열 자료의 상관관계를 조사하기 위하여 Granger Causality를 사용하였다. Granger Causality 테스트 방법은 그림4에 보인 것과 같이 정상상태(stationary state)로 만든 x, y 두 시계열이 있을 때, 식 (1)과 같이 y를 예측하는 선형 모델 보다 식 (2)와 같이 x 시계열을 함께 사용한 선형 모델이 더 정확하다는 것을 입증하는 방식으로 작동한다. 즉 식(1) err1 의 분산보다 식(2) err2의 분산이 유의미하게 적다면 y를 설명 혹은 예측하기 위해 x를 사용하는 것이 유의미할 정도로 정보력을 가진다. 그러나 실제로 x를 포함하여 y에 대한 예측력이 증가하였다고 하여도 두 시계열 사이에 항상 상관관계가 있다고 의미하지는 않는다. 즉 Granger causality 테스트 방법은 관심이 가는 어떤 현상의 예측에 있어서 다른 정보를 활용하는 것이 유의미하다는 것을 보여주는 척도일 뿐이다.

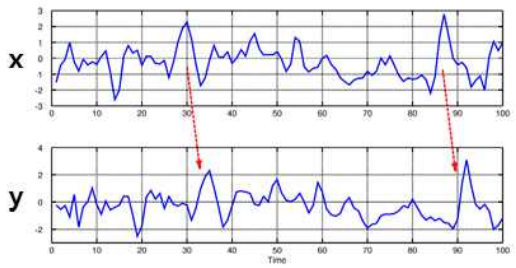

Fig 4. Concept of Granger causality test
그림 4. Granger Causality 테스트 개념

$$y_t = a_0 + ay_{t-1} + a_2y_{t-2} + \cdots + a_my_{t-m} + err_1 \quad -- (1)$$

$$y_t = a_0 + a_1y_{t-1} + a_2y_{t-2} + \cdots a_my_{t-m} + b_pX_{t-p} + \cdots bqX_{t-q} + err_2 \quad -- (2)$$

그림1 의 (c) Granger Causality 테스트의 결과로 출력되는 p-value는 영국의 통계학자 Ronald Fisher가 다른 통계학자들과 정한 기준으로 실험에서 약 5%의 에러를 인정하는 규칙을 만든 것이다[14]. 본 논문에서는 Granger Causality 테스트를 실시하여 p-value 가 0.05 이하인 경우를 유의미한 경우로 정하였다. 이때 x 시계열의 lap 값을 조정하면 x와의 시차를 조정할 수 있고, p-value 가 0.05 이하로 되는 lap 값을 찾으면 x와 y간의 유의미한 상관관계가 있다고 할 수 있다.

식 (3)은 x와 y간의 교차상관계수를 계산하는 식이고 교차상관계수를 최대로 만드는 k의 부호에 따라 선.후행 관계가 성립된다. k < 0 일 경우 y가 x에 선행하는 것을 의미하고 k = 0 인 경우는 동행, k > 0 인 경우는 x가 선행하는 것을 의미한다.

$$r_{xy}(\lambda) = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(y_{t\pm k} - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2(yt - \bar{y})^2}}, where\ \lambda = 0, \pm1, \pm2 \quad -- (3)$$

그림5 는 한국인 출국자 예측을 위해 사용한 자료를 그래프로 표현한 것이다. 한국인 출국자는 일단위 자료이다. 네이버의 트렌드 자료는 주간단위로 제공되며, 비율로 제공된다. 따라서 한국인 출국자 자료를 주간단위로 변환하여 활용하였다.

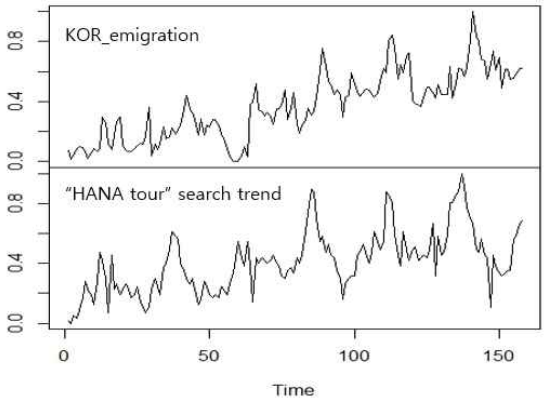

Fig. 5 (upper) Graph for Korean emigration with weekly (lower) Graph for Naver search trend of "HANA Tour"
그림 5. (위) 주단위로 표시한 한국인 출국자
(아래) "하나투어"에 대한 네이버 트렌드

그림5의 자료를 활용하여 Granger Causality 테스트 결과는 그림6, 표1과 같다. 그림6의 하단의 선은 p-value 0.05를 의미한다. 테스트 결과 실제 출국자와 네이버 검색 트렌드 "하나투어"와는 6주 정도의 차이를 갖고 상관관계가 나타남을 알 수 있다.

Table 1 KOR-HANA Granger Causality Test result
표 1. KOR-HANA 에 대한 Granger Causality 결과 표

lap	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	...	-49	-50
p_value	0.7051	0.2216	0.4154	0.2000	0.2512	0.0065	0.0002	0.0001	0.0002	0.0003	0.0009	...	0.7964	0.7799

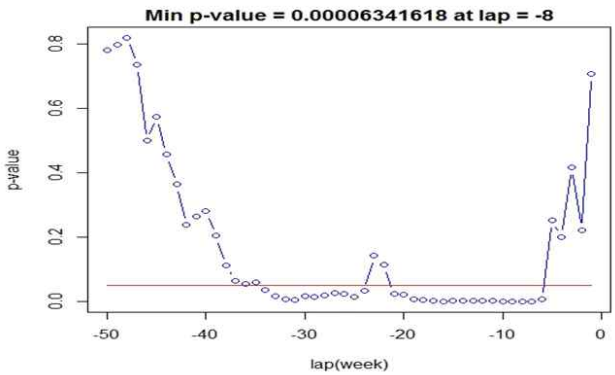

Fig. 6 Graph of KOR-HANA Granger Causality test result
그림 6. KOR-HANA 에 대한 Granger Causality 결과

5. 출국자 예측 모델링

인천공항을 통한 출국자의 예측 모델링을 위하여 선형회귀분석(linear regression), 신경망(neural network) 예측 기법을 활용하였다. 그리고 각 예측

모델링 기법에 “출국자 수”를 활용한 예측과 “출국자 수” + “트렌드 자료“를 활용한 예측 모델링 하는 두 가지 방법을 사용하여 각 모델링 결과를 비교 하였다. ”가“와 ”나”절에서는 예측 관심 날짜의 1일 후의 예측을 수행하여 예측 모델에 적절함을 보였고, “다”절에서는 앞에서 세운 모델을 바탕으로 단기 예측을 실시하였다.

가. 선형회귀분석(linear regression) 예측 모델링

선형회귀분석은 각 입력 변수에 대한 가중치를 least squares error 알고리즘으로 구한 후, 이 선형 방정식을 이용하여 예측하는 것이다. 식(4)에서 주어진 점에서 식(5)의 x(t)를 예측하기 위한 학습으로 x(t-1), x(t-2), …, x(t-n) 을 독립변수(입력변수)로 취해 식(5)의 $(x(t) - x(t-1))^2$ 값을 최소로 하는 βi 값을 구한다 (least squares error 알고리즘). 선형회귀분석은 비선형성질이 강한 시계열 예측에선 상당히 큰 에러를 발생시킨다. 최근에는 선형 회귀 분석보다 신경망을 활용한 시계열 예측이 주류를 이루고 있다. 그렇지만, 데이터 특성이나 목적에 따라 선형회귀분석도 아주 유용할 수 있다.

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \quad \text{-- (4)}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{-- (5)}$$

표2는 한국인 출국자 수를 사용해서 관심이 있는 날의 출국자를 예측한 표이다. 표에서 “KOR″는 인천공항을 통해 출국한 한국인 수 이다. ″n″은 모델링을 위하여 입력한 변수의 개수를 의미하며, n=3 일 경우 예측 날짜 보다 -1, -2, -3 날짜의 자료를 의미한다. 예측 모델링 결과의 효과는 MAPE(Mean Absolute Percent Error) 로 판단한다. MAPE 값은 10%이내 : 매우 정확, 10-20% 이내 : 정확, 20-50% 이내 : 보통, 50% 이상 : 부정확으로 간주한다[15].

선형회귀 모델링은 R언어의 lm 패키지를 사용하였다. “KOR″ 자료만을 사용하여 선형회귀 모델링은 lm_1으로 ″KOR″ 자료와 ”하나투어“ 자료를 사용한 선형회귀 모델링은 lm_2로 명명하였다.

Table 2. Prediction Modeling for the days after using Linear Regression using only ″KOR″ data

표 2. 선형회귀 예측 - “KOR″자료 사용, 하루 후

Forecast lag	n	lm_1 MAPE(%)
1	3	7.64
	4	9.39
	5	9.51
	6	9.67
	7	8.52

표3은 ″KOR″ 자료와 네이버 트렌드의 “하나투어” 자료를 예측 기재로 활용하여 선형회귀 예측 모델링을 수행한 표이다.

Table 3. Prediction Modeling for the days after using Linear Regression using ″KOR″ + ″HANA tour″

표 3. 선형회귀 예측 - “KOR″ + ″하나투어”, 하루 후

Forecast lag	n	lm_2 MAPE(%)
1	3	7.83
	4	7.68
	5	7.77
	6	7.58
	7	6.60

그림7은 두 개의 선형회귀 모델링 결과를 비교한 그림이다. lm_1 보다 lm_2의 MAPE가 적은 것을 볼 수 있다. 이것은 “KOR″ 자료를 독립적으로 사용한 예측결과 보다 ”KOR″ 자료와 “하나투어” 자료를 같이 사용했을 경우에 예측 결과가 더 좋다는 것을 의미한다. 즉 검색 트렌드가 예측 기재로서 훌륭히 작동한다는 것을 의미한다.

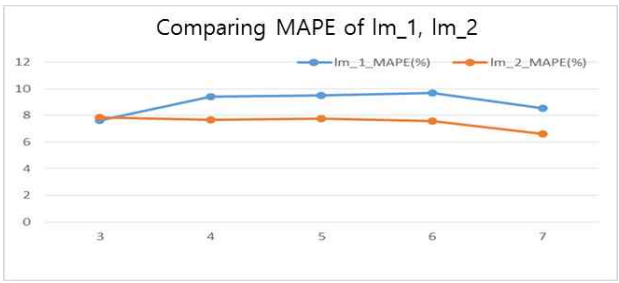

Fig. 7 Comparing MAPE of lm_1, lm_2

그림 7. lm_1, lm_2의 MAPE 비교

나. 신경망(Neural Network) 예측 모델링

신경망 분석은 뇌에서 영감을 얻은 통계학적 학습 알고리즘이다. 신경망은 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을

가지는 모델 전반을 가리킨다[16].

본 논문에서는 신경망 예측 모델링을 위해 R의 nnet과 caret 패키지를 활용하였다. nnet은 Feed Forward Neural Network 즉, 데이터가 앞으로만 흐르는 신경망 초기 모델이다. 학습 알고리즘은 Back-Propagation방법을 사용한다. Back-Propagation은 신경망을 학습시키는 알고리즘 중의 하나이다. 뉴런의 weight를 한 번에 최적화할 수 없기 때문에 식(6)에 제시한 에러 함수 E의 값을 weight 로 미분하여 최소로 만드는 값을 구한다.

$$E = \frac{1}{2}\sum_{l=1}^{L}\sum_{h=1}^{H}(o_{lh} - y_{lh})^2 \quad -- (6)$$

표4는 “KOR” 자료만을 사용해서 신경망 예측 모델링을 수행한 결과이다. “KOR” 자료만을 사용하여 선형회귀 모델링은 nnet_1으로 ”KOR” 자료와 ”하나투어“ 자료를 사용한 선형회귀 모델링은 nnet_2로 명명하였다.

Table 4. Prediction Modeling for the days after using Neural net using only ”KOR” data
표 4. 신경망 예측 - “KOR”자료 사용, 하루 후

Forecast lag	n	nnet_1 MAPE(%)
1	3	7.68
	4	7.58
	5	6.80
	6	7.39
	7	5.36

표5는 ”KOR” 자료와 네이버 트렌드의 “하나투어” 자료를 예측 기재로 활용하여 예측 모델링을 수행한 표이다.

Table 5. Prediction Modeling for the days after Neural net using ”KOR” + ”HANA tour”
표 5. 신경망 예측 - “KOR” + ”하나투어”, 하루 후

Forecast lag	n	nnet_2 MAPE(%)
1	3	7.34
	4	6.98
	5	7.69
	6	7.49
	7	6.65

그림8은 두 개의 신경망 모델링 결과를 비교한 그림이다. 선형회귀 예측 모델링과 달리 nnet_1과 nnet_2의 MAPE가 n에 따라 교차하는 것을 볼 수 있다.

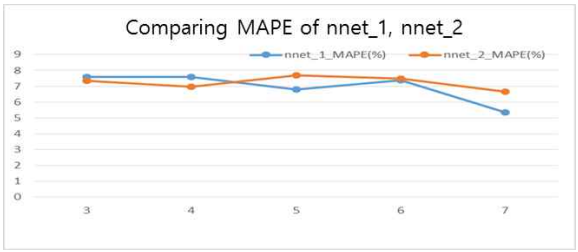

Fig. 8 Comparing MAPE of nnet_1, nnet_2
그림 8. nnet_1, nnet_2의 MAPE 비교

다. 단기 예측

앞 가. 나. 절에서 세운 모델을 활용하여 최대 200일 이후의 단기 예측을 하였다. 표6과 표7은 단기 예측 모델링 후 의미 있는 날짜의 자료이다. 표에서 나타난 것과 같이 단기 예측에서는 “KOR” 단독으로 예측하는 것보다 ”KOR” + “하나투어” 자료로 예측하는 것이 예측률이 더 높은 것으로 나타났다. 그림9는 선형회귀분석을 이용한 단기 예측 그래프이고 그림10은 신경망을 활용한 단기 예측 그래프이다. “하나투어” 예측기재가 10 ~ 75 정도까지는 잡음으로 인식되다가 80일 이후부터 제대로 효과를 발휘하는 것을 확인할 수 있다. 이는 앞서 Granger Causality 분석 결과처럼 단기(8주 이상)에서 예측 기재가 작동함을 입증하는 결과이다. 그림9와 그림10의 화살표는 “하나투어” 예측 기재가 최대로 효과를 발휘하는 부분이다.

Table 6. Prediction Modeling for short term Linear Regression using ”KOR” data only
표 6. 선형회귀 예측 - “KOR”자료 사용, 단기

Forecast lag	n	lm_1 MAPE(%)	lm_2 MAPE(%)
84	6	7.69	6.38
181	5	6.98	5.20

Table 7. Prediction Modeling for short term Neural net using ”KOR” + ”HANA tour”
표 7. 신경망 예측 - “KOR” + ”하나투어”, 단기

Forecast lag	n	nnet_1 MAPE(%)	nnet_2 MAPE(%)
84	5	10.76	7.13
182	3	5.18	5.14

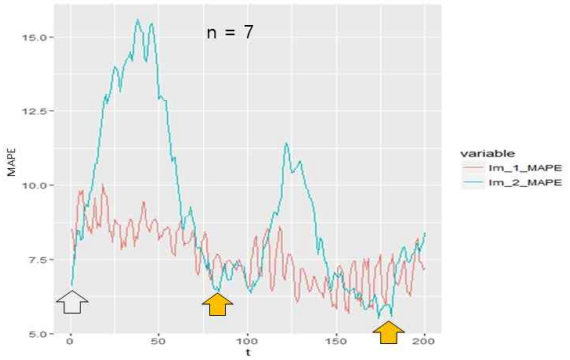

Fig. 9 Forecasting Graph for short term using Linear Regression
그림 9. 선형회귀 예측 - 단기

라. 중국인 출국자 예측

앞 절의 연구 과정을 중국인 출국자(CHN)의 경우로 구현하였다. 중국인 출국자 예측의 정확도를 높이기 위하여 사용한 중국 포탈 "바이두"의 키워드는 "서울"이다. 그림11은 중국인 출국자와 "서울" 키워드 간의 Granger Causality 테스트 결과이다. 약 13주 전에 강한 상관관계를 보인다.

표8은 선형회귀분석 예측 모델링과 신경망 예측 모델링을 "CHN" 단독, "CHN" + "서울"을 각각 수행한 결과이다. 선형회귀 분석의 경우 n 이 증가함에 따라 정확도가 증가하는 모양을 보이나, 신경망 예측의 경우 큰 차이가 없음을 알 수 있다. 그리고 "CHN" 단독의 경우와 "CHN" + "서울"의 경우도 큰 차이가 없었다.

그러나 단기 예측의 경우 표9와 표10의 경우 두 가지 예측 방법론 모두에서 "CHN" 단독 예측 보다 "CHN" + "서울" 자료를 활용한 예측 모델링에서 좀 더 높은 정확도를 보였다. 중국인 출국자 예측에서도 포탈의 키워드 트렌드가 예측기재가 됨을 알 수 있다.

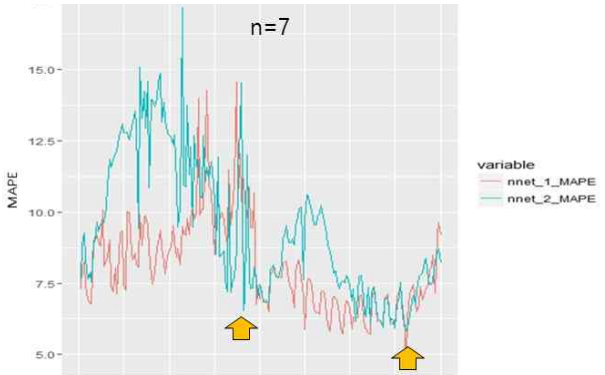

Fig. 10 Forecasting Graph for short term using Neural net
그림 10. 신경망 예측 - 단기

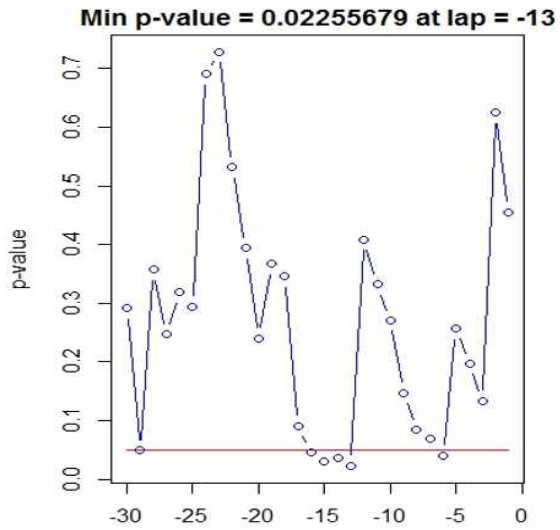

Fig. 11 Graph of CHN-Seoul Granger Causality test result
그림 11. CHN-서울 에 대한 Granger Causality 결과

Table 8. Prediction Modeling for the days after in case of "CHN" and "Seoul"
표 8. 예측 결과 - 하루 후

forecasting lag	n	lm_1_ MAPE(%)	lm_2_ MAPE(%)	nnet_1_ MAPE(%)	nnet_1_ MAPE(%)
1	3	11.55	11.28	11.53	11.31
	4	11.75	11.61	11.26	11.17
	5	11.42	11.33	11.68	11.22
	6	10.82	10.82	10.98	10.65
	7	**10.52**	**10.68**	**10.99**	**10.61**

Table 9. Prediction Modeling for short term Linear Regression using "CHN" data only
표 9. 선형회귀 예측 - "CHN"자료 사용, 단기

Forecast lag	n	lm_1 MAPE(%)	lm_2 MAPE(%)
63	3	11.29	9.13
63	4	11.11	**9.09**
63	5	10.96	9.17
62	6	**10.79**	9.35
61	7	10.88	9.86

Table 10. Prediction Modeling for short term Neural net using "CHN" + "Seoul"
표 10. 신경망 예측 - "CHN" + "서울", 단기

Forecast lag	n	nnet_1 MAPE(%)	nnet_2 MAPE(%)
44	3	12.72	11.08
60	4	12.08	10.62
60	5	12.08	10.62
60	6	**11.42**	**10.23**
60	7	**11.42**	**10.23**

Ⅳ. 실험결과 고찰

인천공항을 통한 출입국자를 예측하기 위하여 한국인 출국자 수를 활용한 단기 예측을 실시하였다. 좀 더 정확한 예측을 위해 포탈에서 여행자가 출국 전에 여행에 관련된 정보를 확인한다는 것에 착안하여 포탈에서 여행관련 키워드에 관한 트렌드 자료를 확보하여 예측기재로 활용하였다.

실험에는 선형회귀분석 예측 모델링과 비선형 예측 모델링인 신경망 예측 모델링을 사용하여 초단기 예측과 단기 예측을 실시하였다. 예측 결과 대부분 사용한 모델링 기법 모두에서 20% 이내의 MAPE를 달성할 수 있었고, 한국인 출국자 예측의 경우 대부분의 구간에서 10% 이내의 MAPE를 보여 매우 정확한 예측이 됨을 보였다. 또 중국인 출국자 예측에서는 10%를 내외의 예측 결과를 보였다. 특히 출국자 수 만을 사용한 예측보다 출국자 수와 포탈의 검색 트렌드를 함께 사용했을 때 단기 예측에서 더 높은 예측 결과를 보여 포탈의 검색 트렌드가 예측기재가 될 수 있음을 보였다.

한국인 출국자 수 예측에서, 단기 예측에서는 선형회귀 모델링이 비선형 모델링인 신경망 예측 모델링 보다 좋은 결과를 보였으나, 장기로 감에 따라 비선형 예측 모델링이 더 좋은 결과를 보였다. 중국인 출국자 수 예측에서는 전반적으로 선형회귀 모델링이 근소한 차이의 정확도를 보였다. 이는 특정 기법이 우세한 결과를 보인다는 것보다는 자료의 특성에 따라 우세한 결과를 나타내는 모델링 기법이 있다는 것을 의미한다. 따라서 예측을 위해서는 다양한 모델링 기법을 사용하여 최적의 예측 결과를 찾아내는 것이 필요하다는 것을 의미한다.

Ⅴ. 결론 및 향후 연구과제

인천공항의 출입국자를 예측하기 위하여 한국인 출국자와 중국인 출국자 자료를 바탕으로 예측 모델링을 수행하였다. 포탈의 여행관련 트렌드 자료를 확보하여 출국자 수 예측기재로 활용하였다. 포탈의 여행관련 트렌드 자료가 출국자 수 예측의 정확도를 높이기 위한 예측기재로 활용됨을 보였다. 한국인 출국자 수를 예측하기 위한 키워드 수집을 위해 다양한 방법을 시도하였으나, 시간과 자원 부족으로 임의의 키워드를 선정하여 출국자 자료와 상관관계를 찾는 방법으로 키워드를 수집하였다. 예측에 적합한 키워드를 찾는 방법론을 확보할 필요가 있다. 또 1개의 키워드를 사용하여 예측을 수행하였으나 상관관계가 높은 다수의 키워드를 활용한다면 좀 더 다양한 연구결과가 나올 것으로 예상된다.

예측 모델링을 수행하기 위하여 선형회귀분석 예측 모델링과 비선형 예측 모델링인 신경망 예측 모델링을 활용하였다. 특정 알고리즘이 우세하기보다는 자료에 따라 적합한 알고리즘이 있다는 것을 확인할 수 있었다. 하나의 자료에 대해 다양한 알고리즘을 적용하여 최적의 예측 모델링을 하는 것이 중요하다. 앞으로 SOFNN(Self-Organizing Fuzzy Neural Network)[17][18] 등 예측 성능이 우수하다고 알려진 알고리즘을 사용하여 예측 모델링을 사용할 예정이다.

한국인 출국자 수 그래프에 나타난 엄청난 스파이크 자료는 예측 모델링의 정확도를 크게 떨어뜨리는 요인으로 작용한다. 스파이크 자료를 제거한 예측 모델링과 스파이크 자료만을 모아서 예측 모델링을 할 예정이다. 다양한 예측 결과를 보여줄 것으로 기대한다.

References

[1] A Study on Forecasting the Demand for Air Demand, Report 11-1611000-002646-14, Ministry of Land, Transport and Marine Affairs, Dec. 2012

[2] Y. Kim, “Study on Low Cost Carrier Demand Forecasting Using Seasonal ARIMA Model,” *The Journal of Tourism Research*, Vol.26, No.1, pp.3-25, 2014.

[3] S. Nam, “A Study on the Air Travel Demand Forecasting using Time-Series Model,” Ph.D thesis, Korea Aerospace Univ. 2010

[4] A. Samagaio, M. Wolters, “Comparative analysis of government forecasts for Lisbon airport,” *Journal of Air Transport Management* 16, pp. 213 - 217, 2010
DOI:10.1016/j.jairtraman.2009.09.002

[5] J. Yoon, N. Huh, S. Kim, H. Hur, “A Study on International Passenger and Freight Forecasting Using the Seasonal Multivariate Time Series Models,” *Journal of the Korean Statistical Society*, Vol. 17 pp 473-481, 2010
DOI:10.5351/CKSS.2010.17.3.473
[6] S. Baik, S. KIM "Estimation of Air Travel Demand Models and Elasticities for Jeju-Mainland Domestic Routes," *Journal of Korean Society of Transportation* 26(1), Korean Society of Transportation, pp 51 - 63, 2008
[7] D.T. Duval, A. Schiff, "Effect of air services availability on international visitors to NewZealand," *Journal of Air Transportation Management*. 17 pp. 175 - .180, 2011
DOI:10.1016/j.jairtraman.2010.12.006
[8] T. Grocshe, F. Rothlauf, A. Heinzl, “Gravity models for airline passenger volume estimation,” *Journal of Transportation Management* 13 pp. 175 - .183, 2007
DOI:10.1016/j.jairtraman.2007.02.001
[9] S. Kim, D. Shin, "Forecasting short-term air passenger demand using big data from search engine queries," *Automation in Construction* 70 pp. 98 - 108, 2016
DOI:10.1016/j.autcon.2016.06.009
[10] Johan Bollen, Huina Maoa, Xiaojun Zeng, “Twitter moods predict the stock market,” *Journal of Computational Science* 2 pp. 1 - 8, 2011.DOI:10.1016/j.jocs.2010.12.007
[11]Naver Trend, http://ca.datalab.naver.com/ca/step1.naver
[12] Baidu Index, http://index.baidu.com
[13] Jerome T. Connor, R. Douglas Martin, L. E. Atlas, "Recurrent Neural Networks and Robust Time Series Prediction," IEEE *Transactions on Neural Networks*, Vol. 5, No. 2, March 1994.DOI:10.1109/72.279188
[14] "p-value", https://en.wikipedia.org/wiki/P-value
[15] Y.Choi, “Forecasting Accuracy of Tourism Demand : An Evaluation of Time Series Methods” Ph.D thesis, Kyungkee univ, 1997
[16]"Artificial Neural net," https://en.wikipedia.org/wiki/Artificial_neural_network
[17] Gang Leng, Girijesh Prasad, Thomas Martin McGinnity, "An on-line algorithm for creating self-organizing fuzzy neural networks," *Neural Networks* 17, 477 - 493, 2004.
DOI:10.1016/j.neunet.2004.07.009
[18] Gang Leng, Thomas Martin McGinnity, Girijesh Prasad, "Design for Self-Organizing Fuzzy Neural Networks Based on Genetic Algorithms," *IEEE Transactions on Fuzzy Systems,* Vol. 14, No. 6, December 2006.
DOI:10.1109/TFUZZ.2006.877361

BIOGRAPHY

Euiseob Shin(Member)

1984 : BS degree in Electronics Engineering, Hanyang University.
1986 : MS degree in Electronics Engineering, Hanyang University.
2014~ : PhD degree in Electronics Engineering, Soonchunhyang University.

Dong-Heon Yang(Member)

1994 : BS degree in Electrical Engineering, Soongsil University.
2010 : MS degree in Biz. Adm.., Inha University.
2014 : Ph.d Dept. of Mgmt. Consulting, Hansung Univ.
1997 ~ Incheon International Airport Corp.

Sei Chang Sohn(Member)

1981 : BS degree in Mechanical Engineering, Yonsei University.
1989 : MS degree in Computer Engineering,Yonsei University.
2013 : Ph.d Dept. of Biz. Adm., Korea Aerospace University.
1985 ~ 1995 KEPCO E&C
1995 ~ Incheon International Airport Corp.

MoonHaeng Huh(Member)

1979 : BS degree in Computer S/W Engineering, Soongsil University.
1989 : MS degree in Computer Engineering,Yonsei University.
2003 : Ph.d Dept. of Computer Eng., Choongbuk Univ. Aerospace University.
1980 ~ 1983 ETRI Researcher
1984 ~ 2000 KT Researcher
2004~ Prof. Dept. of Digital Media, Anyang University.

Seokchul Baek(Member)

1982 : BS degree in Physics Edu. Seoul University.
1984 : MS degree in Theoretical Physics, KAIST
1995 Ph.d Dept. of Statistical Physics, KAIST
1985 ~ 1998 Researcher, KT
2012 ~ NIBD Lab CEO
2012 ~ Prof. Dept. of Digital Media, Anyang Univ.