여행 수요 파악 및 항공 노선 전략 연구 : 웹 크롤링 기반 분석 기법

조창현, 유헌창
고려대학교 컴퓨터정보통신대학원 소프트웨어공학과
e-mail: {chogooood, yuhc}@korea.ac.kr

Study of Travel Demand and Air Route Strategy : Web Crawling-based Analysis Technology

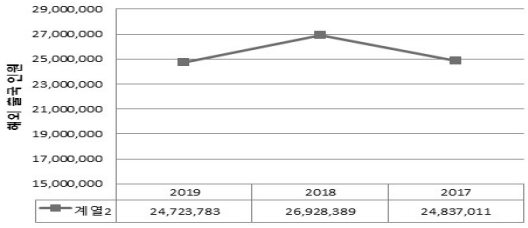
Chang-Hyeon Cho, Heonchang Yu
Dept of Software Engineering, Graduate School, Korea University

요 약

항공/여행 상품은 타 산업보다 불확실성에 취약하며 시간의 절대적인 종속성으로 인해 정확한 수요 파악 및 예측을 하지 못할 경우 가치가 0으로 수렴한다. 이에 본 논문은 웹 크롤링을 기반으로 잠재 여행 욕구를 파악하고, 향후 성장할 것으로 예상되는 항공 노선 및 취항지를 예측 및 분석하는 기법을 제안하고자 한다.

1. 서론

최근 여행 산업의 경우 (그림 1)의 통계 자료에서 보듯이, 2018년 여행자 수는 약 2,600만 명으로 여행업계는 유래 없는 기록을 갱신, 2019년은 조심스럽게 3,000만 명 출국 시대를 기대하는 분위기였다.



(그림 1) 최근 3년간 해외 출국자 수 (출처: 통계청)

하지만 2019년 홍콩 사태, 한/일 관계 악화, 2020년 중국의 코로나-19의 발생 및 확산 공포 등의 예상치 못한 돌발 변수로 인해 그 어느 때 보다 힘들고 혹독한 상황에 놓여 있다. 본 논문에서는 이런 어려움 속에서 여행 수요를 보다 효과적으로 파악할 수 있는 기법에 대해 제안하고자 한다[1].

2. 업계의 불확실성

현재 여행업계가 처한 어려운 상황은 여행업이 타 산업에 비해 불확실성에 취약하기 때문이다. 크게 여행업에서는 2가지의 불확실성이 존재한다. 첫째는 대외상황으로서, 국가 정치적, 질병, 테러, 전쟁 등의 제어하지 못하는 이슈이고, 둘째는 수요와 수익 측면으로서, 고객 수요 파악이 어렵고 여행 상품을 기획하더라도 상품 정형화 및 품질 만족도 제고가 힘들다. 또한, 상품의 판매를 진행하더라도 상황 및 여건에 따라 매출 이익의 폭이 크기 때문에 안정적인 사업 확장이 어렵다. 따라서, 본 논문에서는 두 번째 불확실성에 대한 해결책으로 웹 크롤링을 통해 수요 파악의 불확실성을 최소화하고 정확한 상품 기획이 가능한 기법을 제시하고 검증한다.

3. 분석 및 실험 방법

3.1 수요 측정을 위한 니즈 파악 및 범위

여행 산업에서 고객 수요 파악 형태는 크게 2가지로 나누어 생각할 수 있다.

<표 1> 적극적 고객 정보 활용군 분석표

분류	적극적 고객 정보 활용군	
	(1) 검색형	(2) 대화형
내용	고객이 자신이 원하는 정보를 찾고, 직접 니즈를 입력하는 형태	질문을 통해 고객의 정보를 얻고, 답을 기반으로 직접적으로 니즈를 이끌어내는 형태
서비스 형태	여행 포털, Planner, 여행 오픈 마켓 등…	텍스트 기반 질문 챗봇, 심리테스트, 의사결정 트리 등…

<표 1>의 경우 고객의 정보를 최대한 많이 사용하여 수요를 파악하는 방법이다. <표 2>는 고객의 정보를 최소로 사용하여 수요를 파악하는 방법이다.

<표 2> 소극적 고객 정보 활용군 분석표

분류	소극적 고객 정보 활용군		
	(3) 테마 정의형	(4) 따라가기형	(5) 랜덤형
내용	3자가 여행 테마를 정하고 고객에게 최소한의 선택을 요구하여, 간접적으로 니즈를 파악하는 형태	기존에 만들어진 사례(여행 계획 등…)를 통해 고객이 직접 선택, 결정하게 만드는 형태	고객의 니즈 파악 없이 단순 랜덤 결과값을 제시하여, 최종 선택을 이끌어내는 형태
서비스 형태	테마 여행 상품, BRB 등…	여행 가이드 북, 셀럽 여행, 성지순례 등…	랜덤 박스, 랜덤 게임, 가챠 게임 등…

(그림 2)는 고객 정보 활용군의 포지셔닝을 보여 준다.

(그림 2) 고객 정보 활용군 포지셔닝

본 논문에서는 적극적 고객 정보 활용군 중 잠재적 여행 고객이 가장 대중적으로 사용되는 (1) 검색형 고객 니즈에 대해 분석하고자 한다.

3.2 데이터 수집 범위 및 방법

고객의 수요를 파악하기 위해 먼저 고객의 취향을 반영하는 키워드(맛집, 가족, 살아보기 등…)를 선정하였다.

<표 3> 실험 언어 및 도구 정리

분류	실험 도구	수준
언어	Python	ver : 3.7.4
브라우저	Chrome	ver : 80.0.3987.122
환경	Visual Studio Code	ver : 1.43.0
통계	Excel	2016

<표 3>의 언어와 도구를 통해 실험을 진행하였고 총 3가지 방법을 통해 데이터를 수집하였다. (1) N사의 개발자 사이트 API를 활용하여 블로그, 카페, 뉴스의 키워드 빈도를 파악, (2) 여행 키워드와 높은 결합도를 보이는 단어들의 패턴 파악, (3) 인스타그램의 게시물 수준 및 해시태그(#) 빈도 파악 등 3가지 방법이다.

(1), (2)의 빈도 파악에 카페, 블로그, 뉴스를 중심으로 데이터 수집을 진행하였고 (그림 3), (그림 4) 는 웹 크롤링을 진행한 Python의 소스 예시이다[2][3].

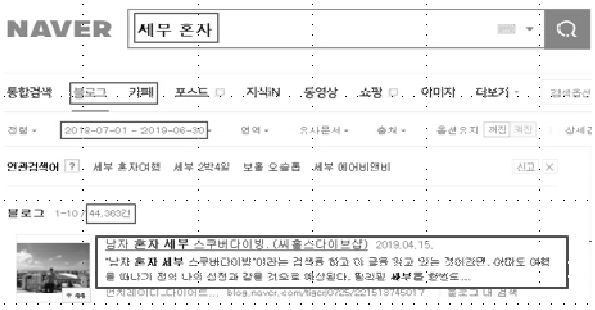
```
import os
import sys
import urllib.request
client_id = "UiYY7dFhUymOLqGO2XGn"
client_secret = "kSivLoybbq"
encText = urllib.parse.quote("치앙마이")
url = "https://openapi.      .com/v1/search/blog?query=" + encText # json 결과
# url = "https://openapi.      .com/v1/search/blog.xml?query=" + encText # xml 결과
request = urllib.request.Request(url)
request.add_header("X-Naver-Client-Id",client_id)
request.add_header("X-Naver-Client-Secret",client_secret)
response = urllib.request.urlopen(request)
rescode = response.getcode()

if(rescode==200):
    response_body = response.read()
    print(response_body.decode('utf-8'))
else:
    print("Error Code:" + rescode)
```

(그림 3) API를 활용한 카페/블로그 웹 크롤링 예시

```
import requests
from bs4 import BeautifulSoup


for page in range(100):
    raw = requests.get('https://search.      .com/search.naver?&wher
    html = BeautifulSoup(raw, 'html.parser')
    articles = html.select('.type01 > li')

    for article in articles:
        journal = article.select_one('span._sp_each_source').text
        title = article.select_one('a._sp_each_title').text

        print(journal, title)
        print('페이지', page, '완료')

print('전체 완료')
```

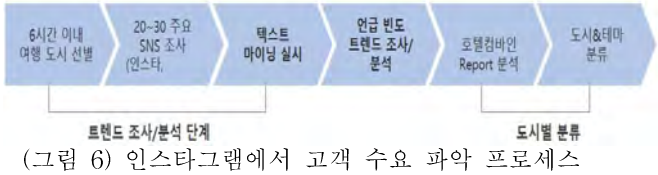
(그림 4) 뉴스 타이틀 및 본문 웹 크롤링 예시

(그림 5)의 경우 N사 중심의 키워드를 어디서 추출했는지를 화면에서 표시한 내용이다. 이 수집에서는 2010년부터 2019년까지의 총 87개 항공편 취항 도시 약 4,190만 건(약 블로그 2,350만개, 카페 1,480만개, 뉴스 360만개)의 여행자 수요 수집을 진행하였다.



(그림 5) N사 포털 실제 데이터 크롤링 영역

(3)의 방법인 (그림 6)의 경우 인스타그램 고객 수요 데이터 수집을 위한 프로세스이며, (1),(2)의 내용과 동일한 총 87개 도시에 대해 총 27,279,336건의 내용을 진행하였다.



(그림 6) 인스타그램에서 고객 수요 파악 프로세스

<표 4>는 87개 도시 중 누적 게시물이 많은 곳을 보여준다. 한국인에게 이미 알려진 휴양 명소, 쇼핑 명소 등이다. 추출 및 분석된 상위 10개의 도시의 경우 한국인에게 성숙된 관광지로 판단할 수 있다.

<표 4> 데이터 분석 기반 성숙 여행 도시

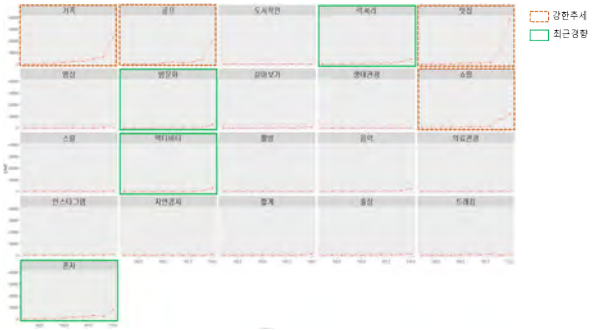
순위	도시	게시물 수	순위	도시	게시물 수
1	세부	5,175,810	6	방콕	1,972,262
2	도쿄	3,492,591	7	괌	1,764,100
3	홍콩	3,476,560	8	다낭	1,318,809
4	오사카	2,438,283	9	후쿠오카	1,243,282
5	상해	2,240,256	10	베이징	977,491

<표 5>는 87개 도시 중 시간(Time)대비 최근 게시물의 증가량이 높은 곳 상위 10개 도시를 추출한 결과이다.

<표 5> 데이터 분석 기반 성장 예상 도시

순위	도시	게시물 증가율	순위	도시	게시물 증가율
1	나트랑	111.60%	6	다카마쓰	47.90%
2	다낭	102.50%	7	하이퐁	45.10%
3	푸꾸옥	92.40%	8	코타키나발루	45.00%
4	가오슝	59.40%	9	타이중	43.90%
5	조호바루	56.50%	10	마쓰야마	42.00%

그리고 분석 당시 수요가 증가할 것으로 예측되는 여행 수요 도시는 나트랑, 다낭, 푸꾸옥 이다. 고객의 수요 내용의 구체화를 위해 각 도시의 자세한 여행 특성을 재분석하였다.

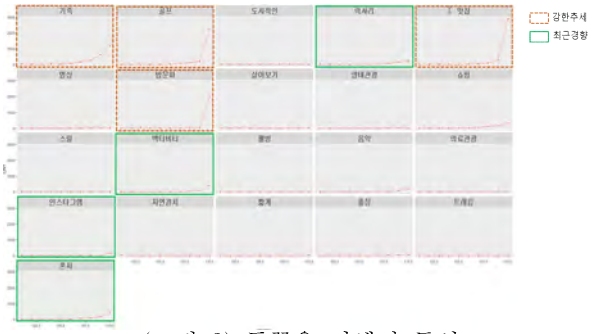


(그림 7) 나트랑 여행지 특성

(그림 7)과 같이 성장 예상 여행지 1위인 나트랑의 경우 기존에는 “럭셔리”, “밤문화”가 중심이었다면, 최근 강하게 보이는 추세는 “골프”, “가족”…등의 키워드이다.

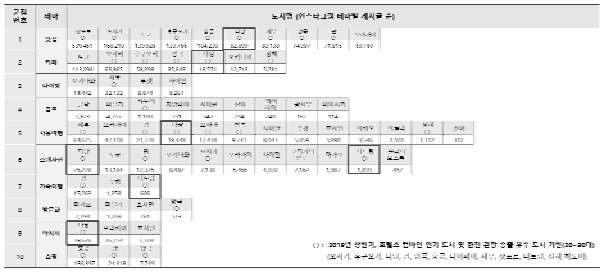
(그림 8) 다낭 여행지 특성

(그림 8)은 2위인 다낭의 여행 특성이며 기존 여행의 테마 및 특성은 “골프”, “혼자”… 여행의 수요가 주를 이뤘고 최근 “가족”, “맛집”, “쇼핑”…에 대한 키워드가 있다.



(그림 9) 푸꾸옥 여행지 특성

(그림 9)과 같이 3위인 푸꾸옥의 경우 기존에 “액티비티”, “혼자” 여행이 주였다면, 최근 “가족”, “골프”, “맛집”의 키워드가 올라오면서 넓은 층의 여행 수요가 잠재해 있다는 것을 예상할 수 있다.



(그림 10) 인스타그램을 통한 고객 수요 파악

(그림 10)의 경우 인스타그램에서 여행자 수요를 게시물과 좋아요 등을 통해 파악된 연관 내용이며, 새로운 여행 수요는 나트랑, 다낭, 푸꾸옥이 추출되며, 연관 여행지 특성으로 “가족여행”, “맛집”, “골프”, “스냅사진” 등이 강한 성장을 보여주고 있다.

4. 분석 내용 검증

4.1 여행지 예상 수요 분석 검증

이번 데이터 수집 및 분석은 2019년도 9월에 진행하였다. 추후 성장 가능할 도시는 나트랑, 다낭, 푸꾸옥으로 예상되었고, 3곳 모두 베트남에 포함된 도시라는 특이점이 있다. <표 6>의 경우 한국인 인기 급상승 해외 여행지에 대한 관련 자료이다[4].

<표 6> 2020년 KLOOK 해외 인기 여행지 순위

순위	도시	성장률
1	**베트남**	**603%**
2	태국	412%
3	인도네시아	260%
4	미국	195%
5	대만	117%

(2019년 1월~12월 예약수 기준 YoY)

<표 7>의 경우 2020년 1월 28일에 만들어진 내용이며, 항공권 검색 및 예약 사이트로 잘 알려진 스카이스캐너(Skyscanner)와 중앙일보의 잠재 여행 고객 설문을 조합하여 발표된 자료이다[5].

<표 7> 2020년 한국인 관심 여행지 Top 10

순위	도시	전년 대비 상승률
1	**푸꾸옥**	**480%**
2	**나뜨랑**	**120%**
3	보라카이	108%
4	치앙마이	57%
5	부다페스트	38%
6	베를린	25%
7	울란바토르	25%
8	블라디보스토크	24%
9	양곤	19%
10	하바나	18%

<표 6>과 <표 7>을 통해 웹 크롤링을 통한 여행 수요 파악이 가능하고 데이터 분석을 통해 향후 성장할 도시를 예측하는 것이 본 검증을 통해 확인되었다[4].

4.2 향후 성장 예상 도시

본 연구에서 도출한 성장이 예상되는 3개 도시의 경우 현재 이미 여행자에게 많이 알려졌고 2020년 상반기에 상품 개발을 적용 시킬 수 있는 여행지이다. 3개의 도시를 제외하고 추가적으로 앞으로 성장이 예상되는 지역에 대해 (그림 11)에서 87개 도시, “추후 성장이 예상되는 군집”을 표시하여 매트릭스에 추가해 보았다.

(그림 11) 신규 성장 예상 도시 매트릭스

지금까지의 가설 및 검증 내용을 적용시키면 동남아지역은 꾸준히 인기가 좋을 것으로 판단되며, 향후 한국인 여행수요가 증가할 것으로 예상되는 국가 및 도시는 “가오슝”, “조호바루”, “다카마쓰”, “하이퐁” 등이 될 것으로 예상 분석된다.

5. 결론 및 향후과제

본 논문에서는 웹 크롤링을 기반으로 잠재 여행 욕구를 파악하고, 향후 성장할 것으로 예상되는 항공 노선 및 취항지를 예측하여 기존의 여행 상품의 가치 불확실성을 줄일 수 있는 효과적인 예측 및 분석 기법을 제시하였다.

먼저 이번 수집에서 N사 개발자 API를 통해 2010년부터 2019년까지의 총 87개 항공편 취항 도시 약 4,190만 건(약 블로그 2,350만개, 카페 1,480만개, 뉴스 360만개)의 여행자 수요 수집하였으며 인스타그램에서 총 27,279,336건의 게시물 및 해시태그(#) 데이터 수집을 진행하였다.

다음으로 이렇게 만들어진 수집 데이터를 기반으로 전체 게시물 수, 시간 대비 상승 수준, 여행 관련 키워드 빈도 등을 고려하여 분석을 진행하였다.

마지막으로 이렇게 만들어진 예측 결과를 실제 여행 상품 판매가 이뤄진 여행 액티비티(KLOOK) 사이트와 항공권 검색(Skyscanner)사이트에서 발표된 2가지 자료를 비교하여 제시한 기법에 대한 검증을 실시하였다.

다만, 2020년의 갑작스러운 코로나-19 이슈로 인해 현재 여행 시장은 물론 한국인의 해외 입국 자체가 불가능하다. 때문에 여행지 수요 예측의 내용을 모두 검증하는 것은 다소 어려우며 현업에 적용시키는 것이 시기적으로 불가능할 수 있다.

추후 시장이 안정화되고 다시 여행객이 증가되는 시기에 실질적으로 현업에 적용하여 제시된 기법을 보완 및 개선해 나갈 필요가 있다.

참고문헌

[1] 조창현. (2019). “항공 예약/발권 시스템의 성능 및 편의성 개선 기법”, 석사학위논문, 고려대학교
[2] 김경수. (2011). “웹 크롤링 수집주기의 동적 설계 및 구현”, 석사학위논문, 충북대학교
[3] 웨스 맥키니, 파이썬 라이브러리를 활용한 데이터 분석 (2013). 한빛미디어, 한국
[4] 파이낸셜뉴스. 베트남 6배 이상 성장... “KLOOK“자료 https://www.fnnews.com/news/202001080919314871
[5] 중앙일보. 2020년 여기가 뜬다… https://news.joins.com/article/236910021