웹 크롤링을 이용한 국가 여행 위험요인 도출 기법†

정다운, 이미현, 유헌창
고려대학교 정보대학 컴퓨터학과
e-mail:{ilikefruit, dnqlalgus, yuhc}@korea.ac.kr

A Deduction Technique of National Travel Risk Factor Using Web Crawling

DaWoon Jeong, MiHyeon Lee, Heonchang Yu
Dept of Computer Science and Engineering, Korea University

요 약

최근 해외로 여행을 떠나는 인구가 증가함으로써 여행의 만족도와 안전성을 높이기 위해 사전 조사의 중요성이 높아지고 있다. 그러나 인터넷을 통해 접근 가능한 SNS와 블로그의 글은 개인이 주관적인 견해를 가지고 작성하기 때문에 신뢰성이 떨어지게 되고 개인이 수집 가능한 정보의 양이 한정적이기 때문에 정확도면에서 한계를 가지게 된다. 본 논문은 웹 크롤링을 통하여 여행 목적지 국가에 관한 뉴스 기사들을 자동으로 수집하고 국가별 위험 요인을 도출하는 기법을 제안한다. 제안하는 기법을 활용할 시 해외여행에 대한 국민의 안전과 만족도가 높아지게 되고 사회 구성원의 전반적인 삶의 질이 향상될 것으로 기대된다.

1. 서론

최근 해외로 여행을 떠나는 인구가 증가하고 있으며, 대부분의 사람들이 자유여행 방식을 선호하고 있다. 이러한 자유여행 방식은 사전 자료조사와 여행일정 설계의 결과에 따라 여행의 만족도와 안전성을 결정한다. 따라서 인터넷의 방대하고 다양한 정보를 분석해서 여행지에 대한 정보를 선별하고 분석하는 작업의 중요성이 높아지고 있다.

일반인들은 여행지 정보를 수집할 시 SNS와 블로그와 같은 주관적인 견해로 만들어진 정보를 접하게 된다. 이러한 정보들은 상업적 목적의 광고로 인해 정보가 오염되면서 신뢰성을 잃어가고 있다. 또한 정보 수집양의 한계로 인해 편협한 의견을 접하게 되면서 정보의 정확도에 한계가 있다.

본 연구에서는 이러한 한계점을 극복하기 위해 객관적인 정보로 인정받은 뉴스 데이터를 대상으로 웹 크롤링을 하여 자동으로 데이터를 수집하고 나라별 여행 위험요인을 분석하는 기법을 제안하고자 한다.

논문의 구성으로 2장에서는 웹 크롤러의 관련연구에 대해서 설명하며 3장에서는 제안하는 국가별 여행 위험요인 분석 기법을 설명한다. 다음으로 4장에서는 제안하는 국가 여행 위험요인 도출 기법의 구현과 실험에 대하여 설명하고 5장에서는 가중치와 관련한 부가적인 실험에 대하여 설명한다. 마지막으로 6장에서는 연구의 결론과 향후 연구 방향에 대해서 설명한다.

† "본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음"(2015-0-00936)

2. 관련 연구

구글 플루 트렌드(Google Flu Trends)는 독감과 연관된 키워드들을 구성하고 키워드들에 대한 구글 이용자들의 검색 트래픽을 분석하여 독감 유행을 예측하는 시스템이다. 이러한 시스템은 2008년부터 2014년까지 진행되었으며, Influenza complication, Cold/flu remedy, General influenza symptoms 등 13개의 세부 주제를 대상으로 질병의 유행 여부를 예측하였다[1].

그러나 구글 검색엔진에서 발생하는 한정된 정보를 기반으로 분석이 진행된 점과 검색 트래픽을 발생시키는 요인이 언론 매체와 같이 다른 종속적인 관계로 인해 영향을 받은 점이 정확도면에서 단점을 보였다.

본 연구에서는 과거에 발생한 사실을 기반으로 작성되는 뉴스 기사를 분석하는 기법을 다루었으며 직접적인 연관관계를 가지는 키워드들만을 도출하여 정확도를 높였다. 그리고 웹 크롤링을 이용하여 방대한 양의 데이터를 수집함과 동시에 정보들의 상대적인 관계를 표현하기 위해 기사 수와 날짜 데이터 기반의 가중치 모델을 설계하였다.

3. 제안하는 국가별 여행 위험요인 도출 기법

이 장에서는 웹 크롤러를 이용하여 국가별 여행 위험요

인을 도출하는 기법을 설명한다. 섹션 1에서는 전반적인 프로세스를 설명하고 섹션 2에서는 수집하기 위하여 선정한 키워드에 대해 설명한다. 이어서 섹션 3에서는 데이터 수집을 하는 방법을 구체적으로 설명한다. 마지막으로 제안하는 기법을 통하여 어떠한 방법으로 위험 요인을 도출하는 지 설명한다.

3.1 제안하는 위험요인 도출 기법의 프로세스

제안하는 도출 기법은 두 개의 대표적인 포털 사이트인 네이버와 다음의 뉴스 기사들을 수집하여 알아보고자 하는 국가의 위험 요인을 분석하여 제공한다. 본 크롤러에서는 클라이언트가 검색을 하고자 하는 나라 또는 지역을 입력 받은 뒤, (키워드 내용) 그 시점부터의 1년까지의 기사를 수집한다. 전반적인 과정은 (그림 1)과 같다.

(그림 1) 분석 과정

3.2 위험요인 키워드

수집하기에 앞서 위험요인과 관련된 대표적인 키워드들을 선정하였다. 키워드들은 크게 세 가지의 범주로 나눠져 있으며, 이는 치안, 자연 재해 그리고 질병이다. 키워드들은 '.csv' 파일로 저장되어 지속적으로 사용하게 된다. 각 분야별 키워드는 <표 1>과 같다.

<표 1> 각 범주 별 키워드

치안	총기, 테러, 전쟁, 마약, 피살, 총격, IS, 폭탄, 실종, 소매치기, 강도, 납치, 감금, 흉기, 폭행, 피습, 방화, 내전
자연재해	태풍 대피, 태풍 강타, 허리케인, 지진, 강진, 산사태, 홍수, 폭설, 가뭄, 산불, 쓰나미, 폭풍, 우박, 화산 폭발, 해일, 사이클론
질병	식중독, 지카바이러스, 메르스, 콜레라, 말라리아, 뇌염, 간염, 에볼라, 독감, 인플루엔자, 수두, 사스, 결핵, 뎅기열, 장티푸스, 홍역, 광견병

3.3 분석 데이터 수집

뉴스 데이터 수집 항목은 포털 사이트의 뉴스 검색 페이지에서 '국가/지역명 + 키워드' 형태의 쿼리를 삽입하였을 때 출력되는 결과에서 '기사 제목', '기사 날짜', '기사 요약' 데이터로 이루어진다. 기사들을 최신 순으로 출력이 되도록 파라미터 값을 설정하고 검색 날짜의 1년 전 날짜까지 나온 기사들만 수집한다.

네이버의 경우는 최대 400페이지, 다음의 경우에는 최대 250페이지까지 열람이 허용되어있기 때문에 최근에 나온 기사의 수가 많은 경우 1년 치 뉴스를 모두 가져 오는 것이 아니라 최대로 가져올 수 있는 기사만큼만 읽어 들인다. 이 중 국가/지역명과 해당 키워드가 제목 또는 요약에 포함되어있는지 여부를 확인하고 출고일자에 따라 가중치를 차등적으로 부여한다. 각 키워드별 가중치는 (1)의 형태로 구성되며, 기호에 대한 설명은 <표 2>에서 보여 준다. 이는 '.csv' 파일 형태로 서버에 저장하게 된다.

$$V_{Nation,keyword} = \sum f_{Nation,keyword,date} \times w_{date} \quad (1)$$

<표 2> 기호 설명

기호	설명
$V_{Nation,keyword}$	가중치 총합
$f_{Nation,keyword,date}$	각 기간별 기사의 빈도수
w_{date}	날짜별 가중치

3.4 위험요인 도출

최종적으로 위험 요인을 도출하기 위하여 두 개의 포털 사이트의 결과 파일들을 합병하는 과정을 진행한다. 각 키워드 별로 가중치 값을 로드하여 합친 후, 새로운 결과 파일을 생성하여 키워드와 가중치 값을 입력한다. 그 후, 값들을 내림차순으로 정렬하여 전체 키워드 중 가장 가중치가 높은 세 개의 키워드를 제공한다. 또한 범주 별로 가중치가 가장 높은 키워드 세 개를 뽑아 출력함으로서 더 자세하게 위험 요인이 무엇인지 알 수 있다.

4. 국가별 위험요인 도출 실험

이 장에서는 제안하는 도출 기법의 실제 데이터 수집 결과를 서술한다. 본 과정에서는 유럽, 아프리카, 아메리카, 아시아 대륙 별 최소 국가/지역 1개 이상을 선정하여 그 나라들의 데이터를 수집하였다. 실험 시 사용한 날짜별 가중치는 <표 3>과 같다.

<표 3> 날짜별 가중치

날짜별 가중치	
당일	100
일주일	70
1개월	50
3개월	30
6개월	10
1년	1

각 국가/지역별로 키워드의 가중치 합을 내림차순으로

정리한 결과, 가장 높은 가중치의 키워드 3개와 각 범주별로 높은 가중치의 키워드 3개를 .csv 파일로 저장하였다. 2017년 9월 9일, 10일 기준으로 얻은 결과는 <표 4>와 같다.

<표 4>에 나온 결과로 보아, 국가/지역 전반적으로는 치안에 대한 위험도가 높은데 반해, 현재 허리케인과 지진의 영향을 받은 아메리카 대륙의 경우 미국과 멕시코의 사례에서 알 수 있듯이 자연재해의 위험도가 높음을 알 수 있다. 또한 최근 테러가 빈번하게 발생하고 있는 유럽에 위치한 영국의 경우에는 '테러' 키워드가 치안의 범주에 있는 다른 키워드들 보다 더 높은 가중치를 갖고 있는 것을 알 수 있다. 뿐만 아니라 각 국가별 유행하는 질병과 가장 많이 발생하는 자연 재해를 확인하여 국가/지역적으로 구별되는 특징이 무엇인지 확인할 수 있다. 사실을 기반으로 한 뉴스를 분석함으로써 검색하고자 하는 나라의 위험 요인들을 정확하게 파악할 수 있다.

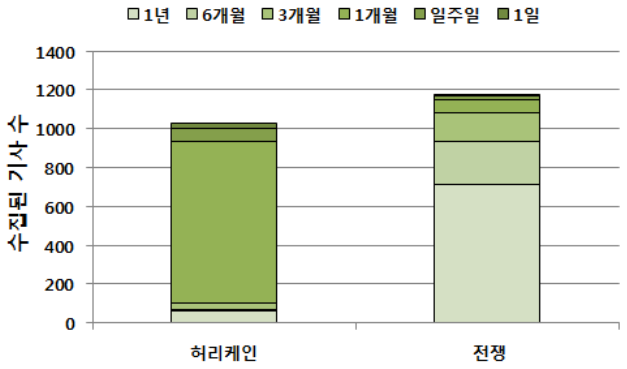
<표 4> 국가별 위험요인 실험결과

국가/지역	전체 Top	범주	Top1	Top2	Top3
나이지리아	테러	치안	테러	폭탄	내전
	폭탄	자연재해	지진	폭풍	홍수
	내전	질병	콜레라	말라리아	간염
남아공	전쟁	치안	전쟁	폭행	테러
	폭행	자연재해	지진	폭풍	가뭄
	테러	질병	식중독	홍역	결핵
도쿄	전쟁	치안	전쟁	폭탄	테러
	지진	자연재해	지진	쓰나미	폭풍
	폭탄	질병	간염	사스	홍역
멕시코	지진	치안	전쟁	마약	강도
	허리케인	자연재해	지진	허리케인	강진
	강진	질병	결핵	식중독	인플루엔자
미국	허리케인	치안	전쟁	폭탄	강도
	전쟁	자연재해	허리케인	지진	폭풍
	폭탄	질병	간염	인플루엔자	식중독
영국	전쟁	치안	전쟁	테러	폭탄
	테러	자연재해	지진	폭풍	홍수
	폭탄	질병	간염	인플루엔자	식중독
인도	전쟁	치안	전쟁	테러	폭탄
	테러	자연재해	홍수	가뭄	지진
	폭탄	질병	말라리아	결핵	간염
필리핀	전쟁	치안	전쟁	IS	테러
	IS	자연재해	지진	홍수	폭풍
	테러	질병	콜레라	자카바이러스	메르스

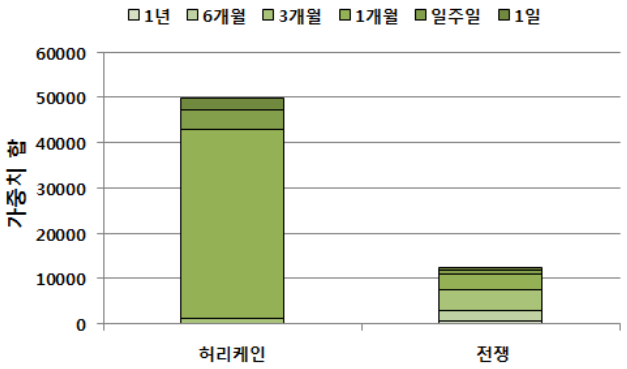
5. 실험

이 장에서는 날짜 가중치와 관련하여 보완하기 위해 진행한 실험에 대해 서술한다. 날짜에 대한 가중치를 부여한 이유는 앞서 언급했듯이 같은 키워드를 포함하는 기사이더라도 최근에 출고된 기사가 더 높은 가치가 있기 때문이다. 각 날짜에 출고된 기사의 빈도를 서로 다른 변수로 설정하고 하루가 지날수록 가중치를 줄이는 linear model을 설계하고자 하였다. 그러나 overfit 현상이 일어나는 것을 방지하기 위해 변수의 수를 줄이고자 <표 3>과 같이 가중치를 임의로 부여하였다.

본 실험은 <표 3>과 같이 설정한 가중치가 위험요인 분석 결과에 어떠한 영향을 미치는지를 보이기 위해 수행되었다. 실험을 위해 사용된 데이터는 실험의 간소화를 위하여 다음에서 수집한 각 기간 별 기사 수 및 가중치를 사용하였다. 실험은 국가별로 두 개의 키워드를 선정하여 두 키워드의 기사 수와 가중치를 누적시킨 결과를 비교하는 방식으로 진행되었다. (그림 2), (그림 3), <표 5> 그리고 <표 6>은 여러 예시 중 멕시코의 허리케인, 전쟁 두 키워드를 비교한 결과이다.



(그림 2) 두 키워드의 기사 수 비교



(그림 3) 두 키워드의 가중치 합 비교

<표 6> 두 키워드의 각 기간별 기사 분포 비율 비교

(단위 : %)

	1일	7일	1개월	3개월	6개월	1년
허리케인	2.4390	6.2439	81.1707	3.5122	0.8780	5.7561
지진	0.3419	1.2821	5.9829	12.8205	18.8034	60.7692

<표 7> 두 키워드의 각 기간별 가중치 비율 비교

(단위 : %)

	1일	7일	1개월	3개월	6개월	1년
허리케인	5.0192	8.9944	83.5190	2.1683	0.1807	0.1185
지진	3.2360	8.4945	28.3149	36.4048	17.7979	5.7520

(그림 2)와 (그림 3)을 통해 키워드 간의 기사 수와 가중치 합을 비교한 결과, 기사 수가 적더라도 전체 가중치가 더 높게 나타날 수도 있다는 점을 확인할 수 있다. 이러한 현상이 나타나는 이유는 <표 5>와 <표 6>을 통해 유추할 수 있다. <표 5>에서 허리케인의 기간별 기사 분포는 지진의 기간별 기사 분포보다 더 현재 날짜에 가깝게 분포하기 때문임을 확인할 수 있다. <표 5>에서 지진의 6개월 전에서 1년 전 기간 안에 출고된 기사의 수가 전체 지진 기사 수의 60%에 육박하는데 반해, <표 6>에서 지진의 동일 기간 기사에 대한 가중치 합의 비율은 단지 5.75%에 불과하다. 이를 통해 가중치가 키워드별 가중치 순위에 영향을 미친다는 점을 확인할 수 있다.

6. 결론 및 향후 연구

본 연구에서 제안하는 국가 여행 위험 요인 도출 기법은 클라이언트가 방문하고자 하는 나라에 대해 사실만을 알 수 있게 하는 지표를 제공함으로서 더 나은 사전 조사를 가능하게 한다. 이러한 기법을 더 확장하여 실시간 서비스를 지원하여 대중에게 제공한다면 공공 데이터로서의 가치도 보일 수 있을 것이라 기대한다. 그러나 두 포털 사이트 모두 보안상의 문제로 특정 페이지 이상 넘어갈 경우 검색이 되지 않는다는 한계점이 있다. 이러한 한계점은 기사 빈도수가 많은 나라와 적은 나라와의 차이를 발생시킨다. 향후 연구에서 이를 개선하기 위해 영국 BBC, 미국 CNN과 같은 해당 국가의 뉴스 사이트 웹 크롤링을 진행하여 지역적인 위험에 대한 데이터를 보강하고자 한다. 또한, 보다 더 정확한 결과를 얻기 위해 날짜 가중치를 조절하고 키워드별 중요도에 따라 가중치를 차등 부여하는 방법도 연구하고자 한다. 그리고 대부분의 키워드는 인과관계가 성립하지만 전쟁과 같은 키워드는 실제 발생한 사건보다는 발생 가능성에 초점을 두고 출고된 기사들이 많기에 직접적인 인과관계가 성립한다고 보기 어렵다. 따라서 이 부분에 대한 개선방안도 마련할 필요가 있다. 마지막으로, 자연어 처리 기술을 통해 문맥적으로 연관성이 떨어지는 기사들을 제거한다면 더욱 높은 정확도를 보일 것으로 기대한다.

참고문헌

[1] Jeremy Ginsberg, Matthew H.Mohebi, Rajan S.Patel, Lynnette Brammer, Mark S.Smolinski, Larry Brilliant “Detecting influenza epidemics using search engine query data” Vol 457, 19 February 2009, doi:10.1038/nature 07634.

[2] David Lazer, Ryan Kennedy, Gary king, Alessandro Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis”, Science, 14 Mar 2014 : 1203-1205.

[3] Declan Butler, “When Google got flu wrong”, Nature 494, 13 Feb 2013 : 155 - 156.

[4] 정동유, 김용태, 박근용, 신재식, 박은주, 임한규, “모바일 웹 크롤링과 GPS를 이용한 지역 뉴스레이터 설계 및 구현”, 한국정보처리학회 춘계학술발표대회 논문집 , 제 24권, 제 1호, 2017.

[5] Castillo, Carlos. "Effective web crawling." ACM SIGIR Forum. Vol.39. No.1., 2005.

[6] Olston, Christopher, and Marc Najork. "Web crawling." Foundations and Trends® in Information Retrieval 4.3, 2010 : 175-246.