ANALYSING TRENDS AND CORRELATIONS FROM INTERNET SEARCHES: CASE STUDY OF ROMANIA

Mioara Popescu[1]

ABSTRACT: *In the recent years, the scientific community explored the possibility to use the publicly available web searches volume histories from search engines together with time-series analysis to forecast and analyse, among other things, travel decisions in specific countries. In this study, we use their approach in a specific case study for Romania, by using query-specific search data to analyse the correlations with other keywords and the seasonality and geographic distribution of the historical online searches. We analyse a data set consisting of searches performed from 2014-2017 using specific keywords. Based on this approach, interesting correlations with different keywords are revealed by those tools. Moreover, we can observe a historical change in the queries performed, which reveal interesting facts about a change of behaviour.*

Introduction and literature review

The increasing volumes of data generated by the online presence and the digital traces captured from the mobile devices reflect various aspects of our activities and represent a new opportunity to study the complex human behaviour [1,2]. The online searches are a prime target source for understanding what people search and when this is happening.

The Internet is very often used for travel planning and the searches performed about the destinations, flight tickets and the time of the year may be useful in predicting where people are more likely to spend their holidays [3]. Nowadays, information regarding almost any subject it is first gathered from online sources. Online searches regarding the holidays reflect the decision taken to choose a time of the year and a location to spend the holidays.

Recently, different search engines have begun to grant access to aggregated data on the number of queries for search terms and the time series of changes generated over time. Google Trends and Google Correlate are some examples of services implemented for historical searches data analysis. In this work, it is presented the investigation regarding the possibility of analysing search query history data from Google Trends and Google Correlate which can provide interesting insights into the information gathering process that precedes the travel decisions recorded in the online searches.

Moreover, together with the travel decisions, other correlation can be found when analysing the queries performed in the same period. In the current paper a specific case study for Romania it is analysed, regarding keywords correlation, timeline and geographically distribution.

Google Trends

Google Trends provides a periodic list of the online search volumes that clients performed on Google search engine previously, on both time series and geographic distribution.

The query index list is obtained by the query share and defined as the query volume for a specific geographic district in time. The most extreme queries in time are normalized with 100, and the query share starting with an initial date is normalized with zero.

Those queries are matched in the sense that queries such as “second-hand cars” are counted in the computation of the query list to “cars”. The database is available starting from 1st of January 2004.

[1] Academia de Studii Economice, Bucureşti, Romania, mioara.popescu@ase.ro

This query list index is accessible at the country level but also at the state/district level, for the United States and several other countries.

Methodology and results

The very first keyword to start the analysis is one of the most used terms when planning a journey is “holidays” (translated in Romanian as “vacante”). Using this keyword as a starting point, Google Correlate returns the list an interesting list of keywords which have a similar query index in time, composed by the most attractive destinations such as Greece, Crete, Jesolo and the names of the most popular travel agencies in Romania. The list of keywords is provided in Figure 1a.

As the first choices are destinations outside of Romania, the most intuitive is that they will also search for the "flight tickets" ("vacante"), which was the second keyword for in this study. Using this keyword as a starting point, Google Correlate returns an unexpected list of keywords which have a similar query index distribution in time. The list of keywords is provided in Figure 1b.

Correlated with **vacante**	
0.8295	sejururi
0.8216	sejur grecia
0.8052	vacanta grecia
0.7968	vacante in grecia
0.7949	creta
0.7933	charter
0.7925	sejur
0.7854	sejur litoral
0.7829	jesolo
0.7812	oferte vacante
0.7804	lido di jesolo
0.7803	vacanta in grecia
0.7803	paralela 45
0.7802	paralia
0.7781	lido
0.7772	turism timisoara
0.7753	pori
0.7750	excursie
0.7729	turism iasi
0.7729	nei pori

Figure 1a.

Correlated with **bilete de avion**	
0.8564	masini second
0.8496	masini second hand
0.8447	harta italia
0.8286	kartago
0.8188	terenuri
0.8186	particulari
0.8162	vanzari case
0.8136	harta spania
0.8130	tunisia
0.8110	credite
0.8103	marshal
0.8065	harta judet
0.8053	etap
0.8053	motors
0.8053	harta auto
0.8047	harta spaniei
0.8009	harta judetului
0.7983	sony map
0.7961	harta
0.7960	distante intre orase

Figure 1b.

Correlation between keywords

Interesting to note that the most similar keyword in term of seasonality index for the “flight tickets” is "second-hand cars" ("masini second hand"). That means that during the observed period, the seasonality of the “flight tickets” was very similar to “second-hand cars”, with a score query index of 0.8564, followed by a similar keyword with a query index of 0.8496. From this interesting correlation, we can conclude that the people searched the flight tickets mainly to travel in other countries for the second-hand car acquisitions.

Looking also at the normalized search activity over time from 2004 to 2017 presented in Figure 2, we can observe that indeed the pattern are very similar over the entire period. We can also observe that there is a peak in 2007, before the crisis from 2008, followed by a decrease in both flight tickets and cars acquisitions, maybe because of the deprecated financial situation.

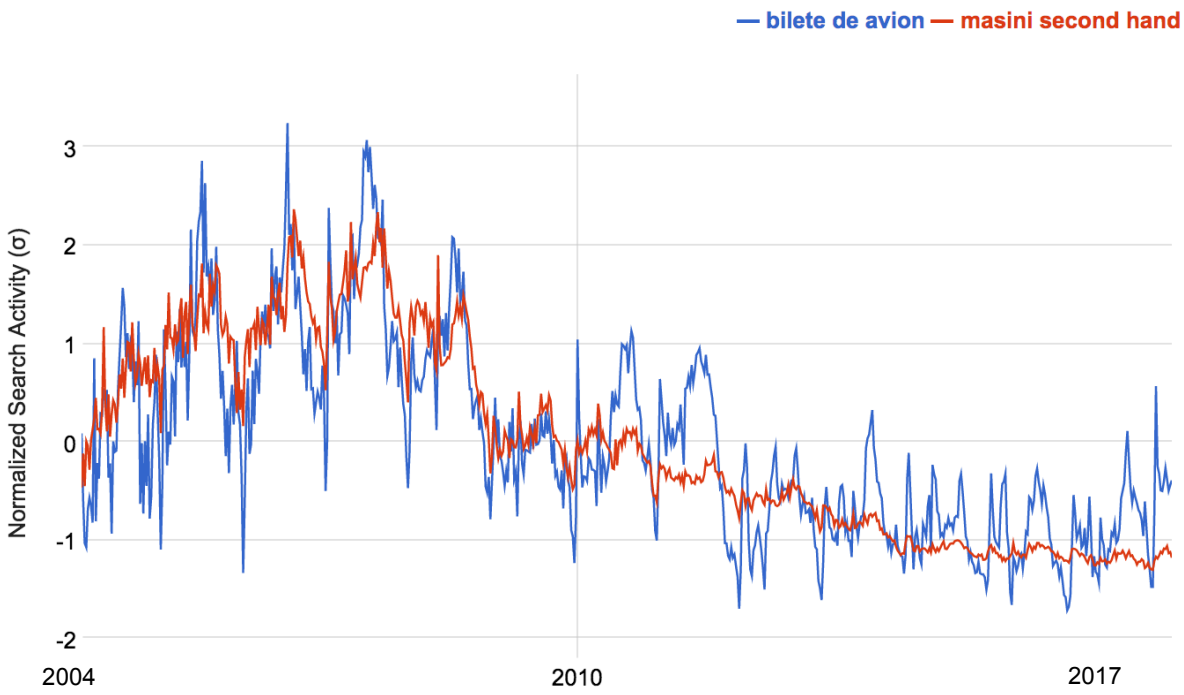

Figure 2. Timeline correlation between two keywords

From the year 2010, the trend looks different and the question that arises, in this case, is related to the factors that influenced the trend disruption, also using the Google Trends service.

Our first hypothesis for the increase of the flight tickets and the decrease of the second-hand car queries was a change of the reason for changing the online search behaviour. Inspired by the Europeans' main purposes of traveling, we analysed the online searches for holidays, in comparison with the flight tickets and second-hand car trends. The results are shown in Figure 3.

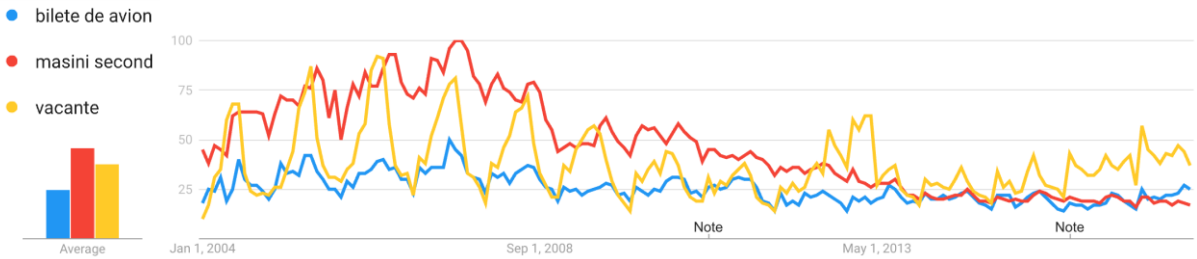

Figure 3. Timeline correlation between three keywords

From Figure 3 can be observed that the volume of search queries for the second-hand cars dropped continuously since 2007 but the online searches for holidays increased. In the same time, the online queries for the flight tickets remained almost constant.

This finding is significant, as can represent the reason for the behavioural change. The conclusion can be that the search for flight tickets in the last years is motivated by the increased trend of traveling for holidays reasons.

Another interesting information can be extracted from the geographic distribution of the online queries. From Figure 4 we can observe that the counties located near the borders and especially in the western part of Romania recorded an increased trend of performing online searches for second-hand cars.

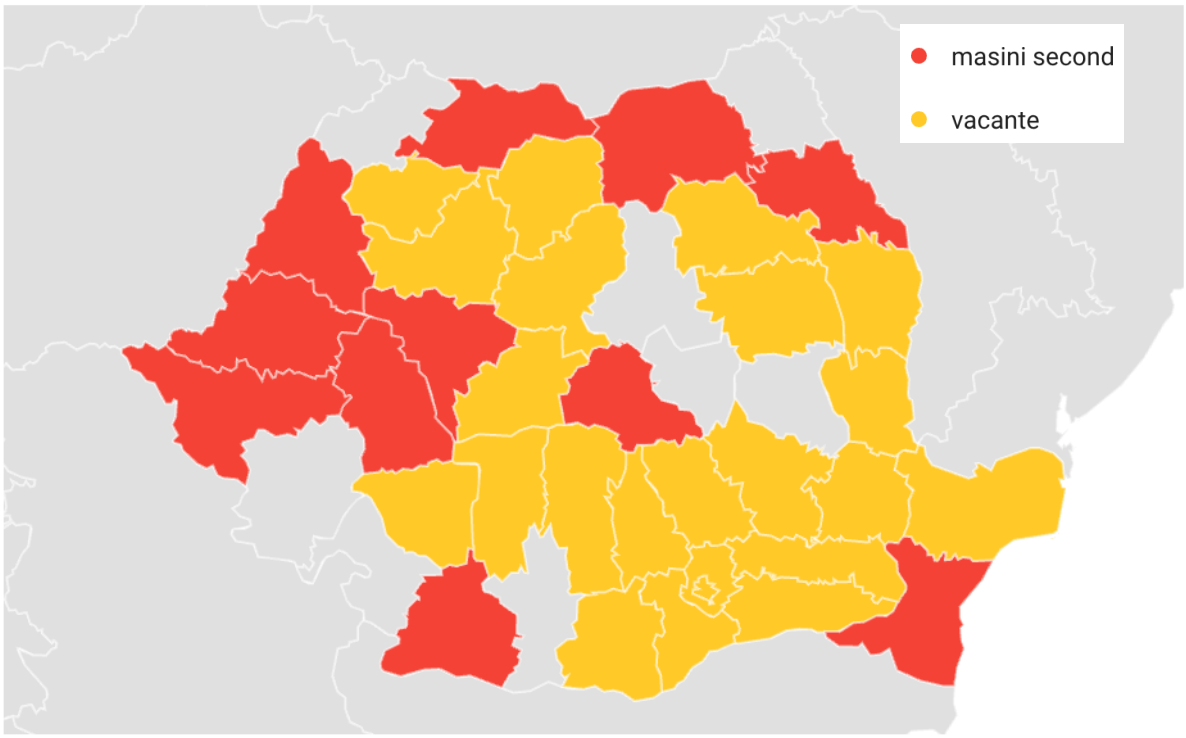

Figure 4. Interest by county for two keywords

This fact is logical as most of the used cars imported in Romania comes from the western part of Europe. Moreover, the counties near the borders hosted important centres of car dealers, because of the proximity of the border and the car registration once imported until the sale for another customer.

Conclusions and future work

In the last years, the online searches proved to be a major indicator of behavioural trends and disruptions in human activity. Interesting facts can be discovered using the correlation between keywords with similar query index, together with the reasons for trends' change in human behaviour. The main advantage is that the trends can be observed at the incipient moment, compared with the classic method of observing the disruptions with a significant delay. This observation means that the solutions for the actual disruptions in the society can be solved proactively rather than reactive.

In the current study, we demonstrate that searching for the correlation of a specific keyword it is possible to find keywords that are correlated. Moreover, the correlations can be from different domains, and the reason can be very unpredictable. The geographical distribution of the online searches can also be an indicator of the trends and people's preferences from different countries and districts.

Further analysis using more keywords and different domains can be done, as an effort to answer to complex questions regarding trends and disruptions. The data mining domain is very complex and the benefits of extracting knowledge from data generated by the human activities can be impressive.

References

1. King, G., 2011. Ensuring the Data-Rich Future of the Social Sciences. Science 331, 719–721.
2. Vespignani, A., 2009. Predicting the Behavior of Techno-Social Systems. Science 325, 425–428.

3. Popescu, M., 2015. Construction of economic indicators using internet searches.
4. Choi, Hyunyoung, Hal Varian, 2012. Predicting the present with Google Trends. Economic Record 88.s1, 2-9.