DATA MINING FOR ANALYSING TRENDS AND CUSTOMER BEHAVIOUR VIA INTERNET SEARCHES

Mioara *POPESCU*[1]

Abstract: The potential of analysis the web searches has been drawing attention of the scientific community for a few years. Despite the large amount of data publicly available from the internet searches, the opportunities for more advanced analysis are still relatively unexplored. The main objective of this paper is to analyse how the web searches can be used as an indicator for different marketing campaigns and strategies. The first analysis explores the possibility to measure the impact and results of the marketing campaigns by the volume of web searches from the period when the campaign was implemented. The second analysis aims to predict the customer behaviour based on the web searches, from the keywords correlation. Public data extracted from search engines has been used for both analyses together with tools used for analysis of trends and correlations of web searches.

Keywords: Data Mining; Web Mining; Correlation; Trends; Pearson Coefficient

JEL Classification: C53, J11, E24

Introduction

The technological progress recorded in recent years in the field of data mining, data analysis, and processing has made possible to exploit an enormous amount of information in any field of economic activity. This fact, correlated with the low cost of storing and processing these large amounts of data, has led to the acceleration of Data Mining research. Because these data processing methods are applied in many areas of economic activity, it is necessary not only an advanced knowledge in Data Mining domain but also a firm theoretical basis complemented by methods practical implementation in each field.

Thus, understanding and exploring the methods of implementing Data Mining practices in various economic sectors and industries is of great importance. Each field has its challenges, and problems that need to be solved, different types of data, but also specific methods that need to be adapted to specific case studies.

Data Mining methods and techniques are used in various economic analyses to capture different economic indicators and used for various economic forecasts (Choi and Varian, 2012). The case studies explored in this paper is the use of Data Mining technologies to build economic indicators using the volume of Internet searches over time (Popescu, 2015). Exploring the web searches have proven to be a quick way to extract valuable information and knowledge of customers behaviour regarding their online activity.

Methodology and results

In this section, we analyse two types of analyses that can be made using the data mining for web searches. The first study will assess a practical example of brand popularity over time, for two

[1] Bucharest University of Economic Studies, 6, Piața Romană, district 1, mio.popescu@yahoo.com

favourite brands. The second analysis deals with the prediction of the customer's behaviour using data mining on volume of web searches over time.

Analyse brand popularity over time using web searches

The online searches can also be used to analyse the brand's popularity over time. Let’s take the example of two trendy brands: Coca-Cola and Pepsi, the manufactures of two of the most popular drinks in the world. From the volume of searches over time, the major events generated by their marketing activities can be analysed. For this example, we use the Google Trends platform to analyse the web searches performed over time, using the Coca-Cola and Pepsi keywords. The web searches volumes from the past five years are shown in Figure 1, using Google Trends tool (Google Trends, 2017).

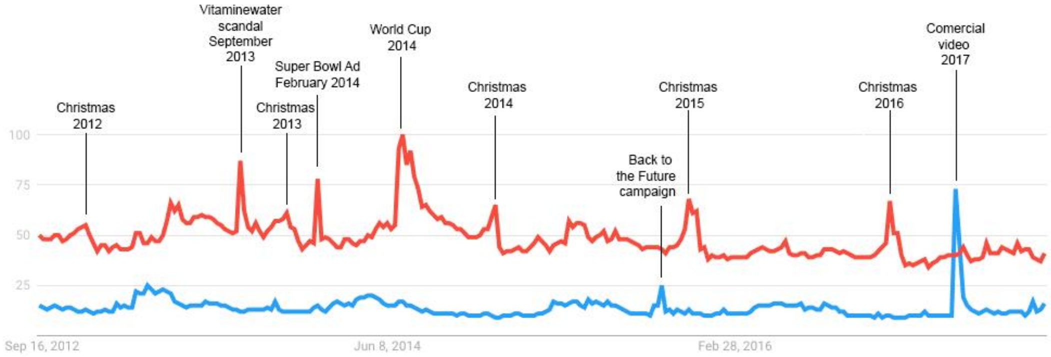


Fig 1. **The volume of searches for the keywords Coca-Cola (red) and Pepsi (blue) in the last five years, with the identification of each high peak**

Source: Google trends

The major peaks represent particular events generated by different marketing activities, strategies and decisions from each brand history. As one of the long-term marketing department objectives is to increase the brand popularity over time, the web searches can be used as a tool to measure the impact and results of those activities and strategies.

Coca-Cola is well known about the marketing investments for the Christmas period. The brand created the viral image of Santa Claus, a funny old man dressed in red. We can see that their campaigns are indeed very efficient, as each year there is a peak of searches during the Christmas time. Apart from the Christmas campaigns, they were active also in another large-scale campaign.

As we can observe from Figure 1, the largest peak can be seen in 2014, when Coca-Cola was the primary sponsor of the World Cup. Even if Coca-Cola reported a bad period of sales compared with the investments in that year, is the official sponsor of the several prestigious soccer competitions around the world, boosted their overall brand image over time.

Another notable marketing event is the video advertisement from February 2014, presented during the Super Bowl, the most prestigious American football competition. That event broke the record for the most watched competition and television show in the history. Surprisingly enough, the advertisement was not received well by the Americans, who considered its multilingual script against their efforts to conserve the language and the country’s tradition. We can conclude that even if the marketing effort fail and people get offended, this creates an effect of buzz news and everyone search to see how bad was the advertisement in this case.

In the previous year, a big scandal arose around a marketing campaign designed to personalise the bottle lids of Vitaminwater, another Coca-Cola product. Because of Canada’s bilingual context, there was a mistake on combining English and French words. As a result, there were some bottles with the message “you retard” with the intention of using the French word “retard” which means “late” in English. The unfortunate one bottle was open by a family who had a daughter who is cognitively delayed. As a result, the story was viral and people around the world shared and search for the details of this story. This is another example where the negative news fired web searches and in the same time brand awareness.

The Pepsi brand is in a similar situation, with an excellent marketing campaign in 2016 and a very uninspired commercial video in 2017. Without a lot of impressive marketing campaigns and buzz news, the notoriety of the brand is lower than their main competitor.

Apart from the volume searches for each brand, Google Trends offer the possibility to explore also the geographic distribution of the searches. In Figure 2 we can see a geographic distribution for both brands, over the last five years. The distribution shows the countries where a particular keyword is dominant over the other.

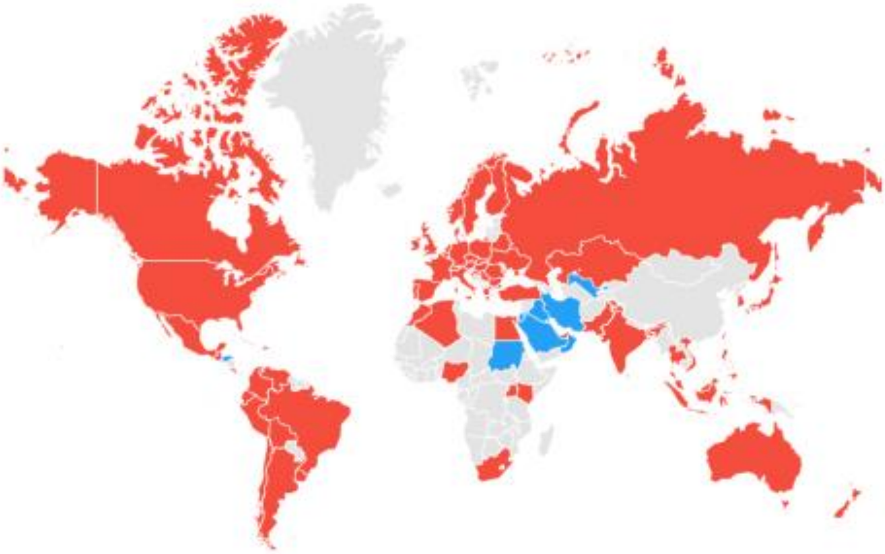


Fig 2. **The geographic distribution of web searches for Coca-Cola (red) and Pepsi (blue) keywords in the last five years**

Source: Google trends

The geographic interest can be analysed not only for an extended period but also for a particular range. As an example, if we take the same geographic distribution from Figure 2 only for the period (2-8 April 2017) when “Pepsi” keyword had the highest volume search peak, we can analyse where people find this news over the world. In Figure 3 we can observe the dominant interest by geographic regions.

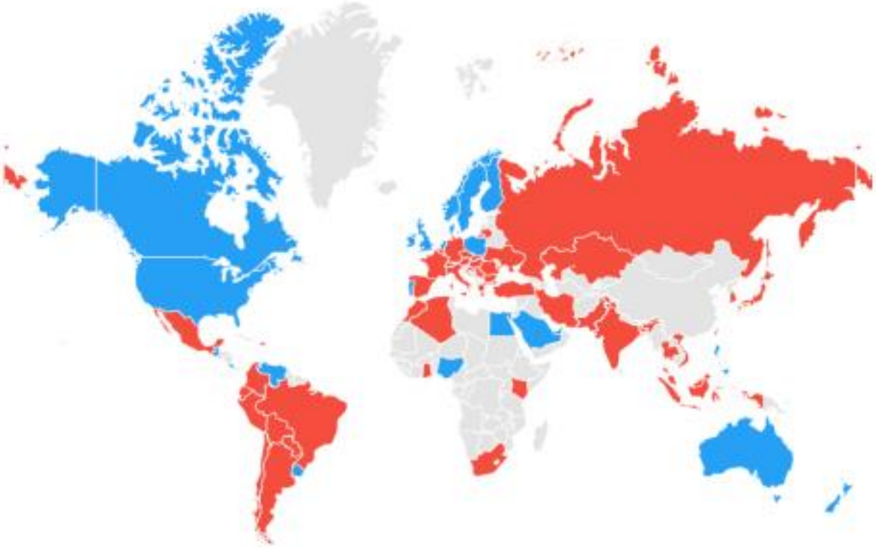


Fig. 3. **The geographic distribution of web searches for Coca-Cola (red) and Pepsi (blue) keywords in the last five years**

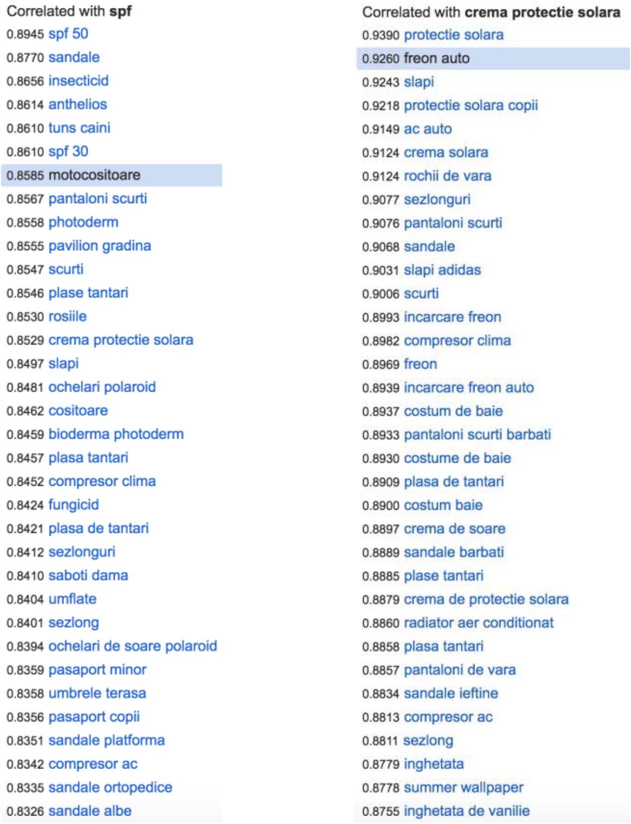
Source: Google trends

From this consideration, we can conclude that the internet searches can be used not only to analyse the volume of searches and trends over time, but also the geographic interest for specific events. It is a great tool for every marketer as the impact and results of each marketing campaign can be measured very fast.

Customers' behaviour prediction using data mining on web searches

A classic example of data mining usage to predict the client's behaviour is the Walmart story of predicting what the customers will buy ahead of hurricanes. It is being one of the examples given by (Walker, 2014), in a series of business cases where the "big data" is used to predict different events and disruptions. Because Walmart records and uses the data from the purchase made by each customer, they tried to predict the top sales products before a hurricane. They found that a particular snack Pop-Tarts was bought in massive quantities before the hurricane. When the National Weather Service stated that hurricane Frances was about to come, they filled the stocks with this product and even to put the tarts near the entrance and in the strategic points so that anyone can see them. The result was an increase of sales in that period.

Our practical example demonstrates how the web queries can offer valuable information regarding the correlation between searches for different products. In our example, we took two keywords for seasonal products to analyse the relationship between the targeted goods and unexpected other products. The chosen keywords were "sunscreen" (crema protectie solara) and "spf" (sun protection factor). The results from Figure 4 present the list with the most correlated keywords in descending order starting with the most correlated ones.



Correlated with **spf**

0.8945	spf 50
0.8770	sandale
0.8656	insecticid
0.8614	anthelios
0.8610	tuns caini
0.8610	spf 30
0.8585	motocositoare
0.8567	pantaloni scurti
0.8558	photoderm
0.8555	pavilion gradina
0.8547	scurti
0.8546	plase tantari
0.8530	rosiile
0.8529	crema protectie solara
0.8497	slapi
0.8481	ochelari polaroid
0.8462	cositoare
0.8459	bioderma photoderm
0.8457	plasa tantari
0.8452	compresor clima
0.8424	fungicid
0.8421	plasa de tantari
0.8412	sezlonguri
0.8410	saboti dama
0.8404	umflate
0.8401	sezlong
0.8394	ochelari de soare polaroid
0.8359	pasaport minor
0.8358	umbrele terasa
0.8356	pasaport copii
0.8351	sandale platforma
0.8342	compresor ac
0.8335	sandale ortopedice
0.8326	sandale albe

Correlated with **crema protectie solara**

0.9390	protectie solara
0.9260	freon auto
0.9243	slapi
0.9218	protectie solara copii
0.9149	ac auto
0.9124	crema solara
0.9124	rochii de vara
0.9077	sezlonguri
0.9076	pantaloni scurti
0.9068	sandale
0.9031	slapi adidas
0.9006	scurti
0.8993	incarcare freon
0.8982	compresor clima
0.8969	freon
0.8939	incarcare freon auto
0.8937	costum de baie
0.8933	pantaloni scurti barbati
0.8930	costume de baie
0.8909	plasa de tantari
0.8900	costum baie
0.8897	crema de soare
0.8889	sandale barbati
0.8885	plase tantari
0.8879	crema de protectie solara
0.8860	radiator aer conditionat
0.8858	plasa tantari
0.8857	pantaloni de vara
0.8834	sandale ieftine
0.8813	compresor ac
0.8811	sezlong
0.8779	inghetata
0.8778	summer wallpaper
0.8755	inghetata de vanilie

Fig. 4. **List of correlations results for the two keywords**

Source: Google trends

In order to analyse the relationship between those keywords and other unexpected products, we used the Google Correlate tool (Google Correlate, 2017). This tool is considered a reverse of Google Trends, as Google Correlate returns other queries with a similar pattern of frequency, synchronised in time. The platform is performing a Pearson Correlation coefficient and returns other keywords with the highest correlation, in the [-1,1] interval.

In the case of the “spf” keyword, except the first six keywords that can be expected in the summer season, another interesting keyword is “motocositoare” (grass cutter machine). A comparison between the correlation over time is presented in Figure 5.

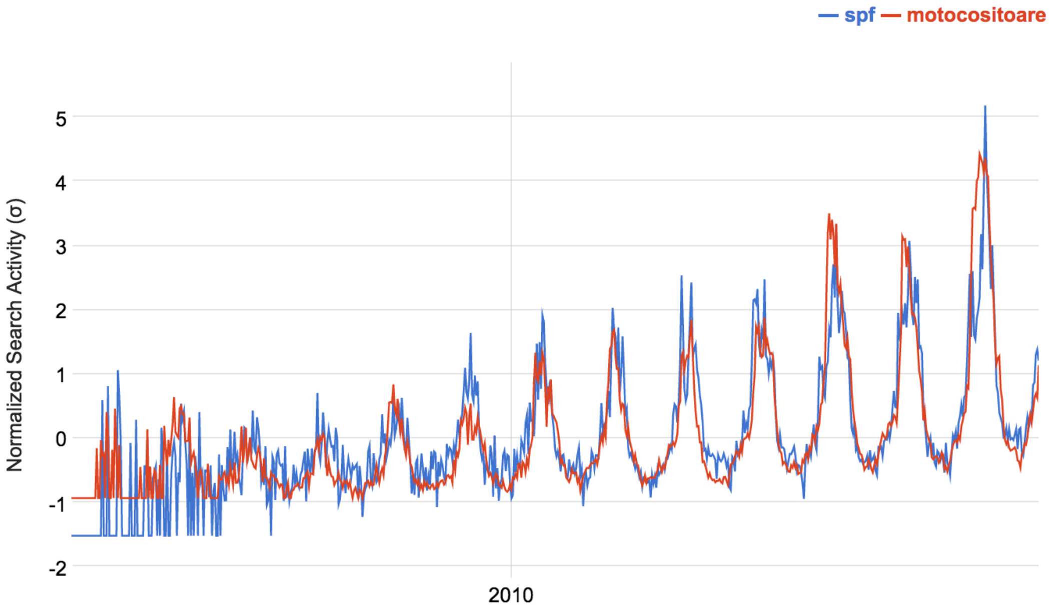


Fig. 5. **Correlation in time between the inspected keywords**

Source: Google trends

The association between the selected keywords is logical but hard to detect without the information from the web searches. The summer season is also the period of mowing the grass. Means that people are searching in this time also to buy the machines or parts and in the same time they search for sunscreens as they also plan to leave on holiday, most probably at the seaside.

Nevertheless, because the correlation is quite consistent over the years, it is clear that this was not just a random search, but it is an identified customer behaviour using the web searches. It can be a valuable information from the marketing perspective as different marketing campaign can be launched to advertise in a way the two products at the same time.

A similar analysis can also be made for the second keyword ("sunscreen"). The results are presented in Figure 6.

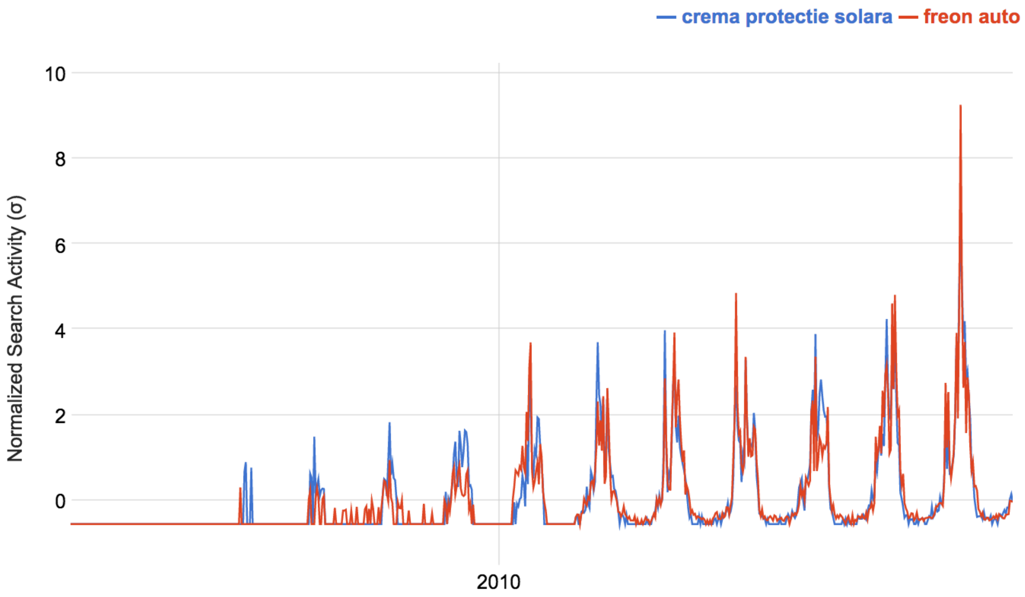


Fig. 6. **Correlation in time between the inspected keywords over time**

Source: Google trends

If the first correlated keyword is enough logic (“sun protection”), the second correlated keyword (“Freon auto”) is very interesting and quite hard to guess without this type of analysis from the web searches. The correlation is indeed logical as in the summertime the temperatures are very high in Romania. People plan to go in holidays, and they must prepare their cars for long trips. In order to be sure that they will have a pleasant journey, they must make a car maintenance, especially the air conditioning system. As a recommendation, they must change each year the cooling agent used in most air conditioning systems.

Conclusions

In this paper, we analyse the possibility to use the data mining on web searches to extract valuable knowledge regarding the impact of marketing campaigns and the customer's behaviour.

The first analysis explores the brand popularity and trends over time, and we demonstrate how the impact of different marketing campaigns can be measured, on both volume of searches and geographical interest. We showed that the incidence of each marketing campaign on both short and long-term could be assessed by the web search analysis.

The second analysis aims to predict what the customers will buy when searching for a particular product. Using Google Correlate and the publicly available data from search engine we demonstrate that especially for seasonal products can be found other exciting products that customers search in the same period. It can be valuable information for retailers as they can manage different marketing campaigns and promotions based on this information.

References

1. Choi, H., Varian, H. 2012. Predicting the present with Google Trends. Economic Record, 88(s1).
2. Google Correlate, 2017. https://www.google.com/trends/correlate
3. Google Trends, 2017. https://www.google.com/trends/
4. John Walker, S. 2014. Big data: A revolution that will transform how we live, work, and think.https://www.google.com/trends/correlate

5. Popescu, M. 2015. Construction of economic indicators using internet searches.