An Algorithm based on Google Trends' data for future prediction. Case study: German Elections

Spyros E. Polykalas
TEI of the Ionian Islands
Dept. of Business Administration,
email: s.polykalas@teiion.gr

George N. Prezerakos
TEI of Piraeus,
Dept. of Electronic Computer Systems,
Petrou Ralli & Thivon 250, Athens,
122 44 Greece,
email: prezerak@teipir.gr

Agisilaos Konidaris
TEI of the Ionian Islands
Dept. of Business Administration,
email: konidari@teiion.gr

***Abstract*— The analysis of the high volume of statistics generated by web search engines worldwide on a daily basis, allow researchers to examine the relation between the user's search preferences and future facts. This analysis can be applied to various areas of society such as sales, epidemics, unemployment and elections. The paper investigates whether prediction of election results is possible by analyzing the behavior of potential voters before the date of the elections. In particular, the proposed algorithm is applied on the three more recent German elections. The results of this analysis show that a strong correlation exists between the search preferences of potential voters before the date of the election race and the actual elections results. It also demonstrates the fact that search preferences are influenced by various social events that may take place concurrently to the election race. The effect of such events has to be filtered out as noise in order to arrive at a successful estimation of the final results.**

Keywords*— *Search Engines Data, Google Trends, Elections, Prediction, Data Mining

I. INTRODUCTION

The analysis of information that is provided by popular search engines (Google, Yahoo etc) with respect to the volume of searches for specific terms can allow the early detection of trends in many areas of social and financial life. In some cases, this detection is so accurate that one may claim that the ability exists to predict the future [1]. One such area is the prediction of election results in national elections in several constituencies.

In this paper an algorithm has been applied to the data provided by the Google Search engine via the Google Trends service in order to examine the relation between the search preferences of web users and the results of the German national elections of 2005, 2009 and 2013. In particular the analysis is focused on the selections of the appropriate set of search terms, in an appropriate timeframe in order to arrive at an accurate estimate of the election results with respect to the two major parties in Germany.

It becomes evident that in there is a strong correlation between the users web-search behavior and the final decision of the voters, a correlation that is so strong that it can lead to the prediction of actual election percentages. The paper is structured as follows: Section 2 discusses other research efforts in the same area. Section 3 describes the algorithm in question. Section 4 is concerned with the three most recent elections in Germany and explains how the algorithm can be for results prediction. Finally, the paper's conclusions reside in Section 6.

II. PREVIOUS WORK

There are several papers that deal with election prediction in the United States such as [2] and [3]. Flickr is discussed as a potential election prediction source in [4] while the same case for Twitter is presented in [5]. On the other hand, arguments against using the web as a means of election prediction are voiced in ([7].

The current paper uses the main principles of our previous work [8] for an algorithm towards election prediction and applies this algorithm on the three more recent national elections of Germany, including the elections of 2013. The paper mainly focuses on the prediction of the election winner, as well as the prediction of the percentages of the two main rival parties. In addition, the impact of several parameters on the results of the algorithm is examined, such as the duration of the pre-elections period used for data collection, the incorporation of historic data from previous election races and the elimination of data-noises generated by events not relevant to the voters' final choice.

The results of the analysis indicate that the aforementioned parameters may influence the accuracy of the predictions, therefore a proper adjustment of the parameters is required taking into account the framework inside which each election race takes place. It should be noted that regardless of the parameters adjustment, the results of the algorithm accurately pinpoint, in all cases, the winner of the elections before the elections date.

III. THE ALGORITHM

Nowadays the high penetration of Internet, allow us to make the safe assumption that a high percentage of citizens in a society are at the same time active internet users . On the other hand the high increase of web searches for words / phrases related to elections, as presented in the following section, during a period before the date of elections, shows that the

978-1-4799-4796-6/13/$31.00 © 2013 IEEE

potential voters use web search as a tool that guides them, to a significant degree, during the election process.

The first question that should be answered is whether internet users and more specifically the users of search engines constitute a representative sample of potential voters. Taking into account the high penetration of Internet and the high usage of search engines, it could be argued that the answer is positive.

Another crucial issue is whether the behavior of users with respect to search can be correlated with the final decision of the potential voters. In other words, does the fact that someone searches on the Internet for a party during the pre-election says something with respect to his/her actual voting preference on the day of the elections? If yes then the volume of such searches would be proportional to the number of votes in the elections.

Of course in reality, the relation between the search term popularity and the final election results may vary due to several factors, such as: the profile of the voters of each party (more vs. less internet friendly), the percentage of participation in each election race, events during the pre-election periods that impact the popularity of search terms but are not necessarily correlated with the final decision of the voters, and more important whether a person that searches for a party, during the pre-election period, will finally vote for that party. In order to eliminate the noise generated by the aforementioned issues the proposed algorithm contains, among other adjustments, logical interventions which deal with the issues mentioned above on a case by case basis.

The proposed algorithm is implemented via the following steps:

A. Selection of the initial set of terms

In the beginning, the initial set of search terms is determined. The two main parties in the German elections are the coalition of the Christian Democratic Union / Christian Social Union (CSU/ CDU) and the Social Democratic Party (SPD). Therefore queries towards Google Trends are initially carried out using the acronyms of the parties as search terms for a few weeks prior each elections day. If a significant variation in search volumes is detected then the search term in question is a valid one for our algorithm.

For each search term, Google Trends returns a normalized averaged number that corresponds to the volume of searches for the specific term (on a daily basis) compared to the rest of the search terms. This number is called the Web search Interest (WI) of each specific search term.

Figure 1 illustrates the Web search Interests between 2004 and 2013. As expected the WIs for parties' acronyms have their highest values around the dates of national elections.

B. Definition of the pre-election periods

In Germany the three last national elections were held on, 18th September of 2005, 28th September of 2009 and 22th September of 2013. As depicted in Figures 1 and 2, the web search popularity before the date of elections presents significant variations in the period approx. 30 days before the election races. Therefore we choose the period of 30 days before the election date as our observation window.

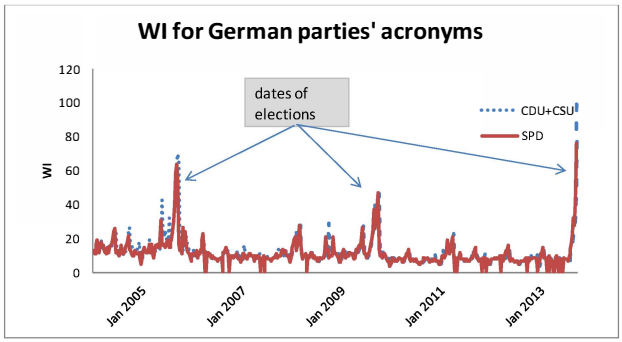

Fig. 1. Web Interest for parties' acronyms

C. Finalization of the set of search terms

At this stage more terms are examined for possible relevance to the elections, aiming to include all search terms that may be related to voters' willing to vote for a specific party in the forthcoming national elections. In order to examine whether a specific word or phrase should be included in the set of search terms for the relevant party, the following two rules are applied: first we examine whether the variation of web interest has peak values around election days, and secondly we examine whether WI values have a variance comparable to the variance of the respective WI for the party's acronyms.

If both criteria are fulfilled then the word/phrase should be included in the set of search terms selected for a specific party. As expected due to the person-centric atmosphere of the German elections, applying the above rules in several runs, we arrive to the conclusion that the names of the parties' leaders have significant impact on the relevant Web Interest and in some cases their impact is significantly higher than the relevant impact of the parties' acronyms.

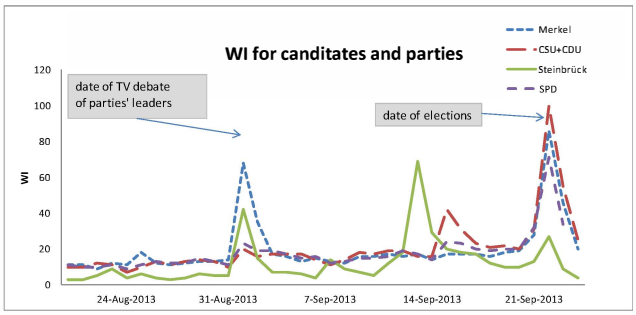

Fig. 2. WI for parties' acronyms and leaders names on 2013 Elections

Finally we have excluded from our input data set, the search volumes corresponding to users searching for either parties or both leaders at the same time. It is assumed that the respective data cannot be used to determine the user's voting preferences therefore it is filtered out as noise.

Table 1 contains the final set of search words for each election race.

Final Set of Search Words			
	2005	*2009*	*2013*
CSU / CDU	CSU + CDU - SPD	CSU + CDU – SPD	CSU + CDU - SPD
	Merkel - Schröder	Merkel - Steinmeier	Merkel - Steinbrück
SPD	SPD - CSU - CDU	SPD - CSU – CDU	SPD - CSU - CDU
	Schröder - Merkel	Steinmeier – Merkel	Steinbrück - Merkel

Table 1. Final set of search words

Figure 2 depicts the Web Interest values of the selected terms for the 2013 elections, while Figure 3 and Figure 4 present the relevant WI values for the 2009 and 2005 national elections respectively.

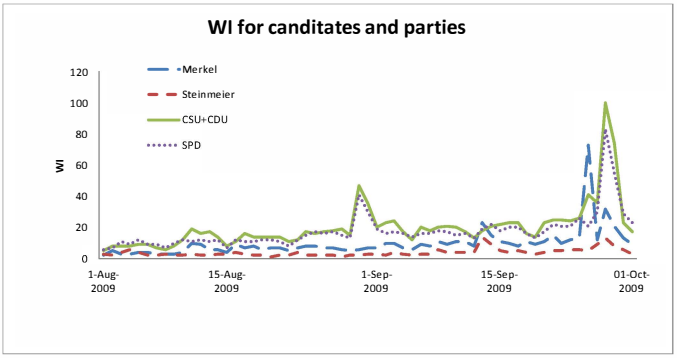

Fig. 3. WI for parties' acronyms and leaders names on 2009 Elections

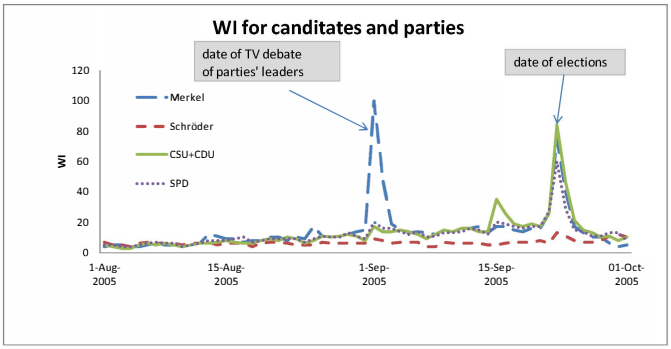

Fig. 4. WI for parties' acronyms and leaders names on 2005 Elections

Further removal of noise from the initial input data set requires that we take into account other factors that modify the WI but are not indicative of the users' willingness to vote for either party.. It is obvious from Figure 2 that the WI of the two parties have almost the same variation or follow the same tendencies driven by several factors during the pre election periods. For example on Sept. 1st 2013, a debate was held between the leaders of the two main parties. This fact influences the WI of the two parties as depicted in Figure 2.

On the other hand, around the 12th of September, the WI of the word "Steinbrück" presents high variation that is not followed by a similar variation of the WI for the name of the leader of the competitor party. On that day, the leader of the SPD gave an interview to the national ARD TV channel and that was the main reason for the significant variation of the relevant WI.

This event did not include a representative of the rival party therefore the significant increase of the WI for "Steinbrück" does not necessarily reflect the willingness of potential voters, who seem to search for this term more out of curiosity about what happened in the specific event and less because they really want to vote for the relevant party. Therefore an adjustment is required to the input data set of the algorithm in order to eliminate the noise generated by this event. According to this logic, the WI values for the name of that leader around the TV show dates are ignored. Figure 5 shows the actual WI and the WI after the adjustment.

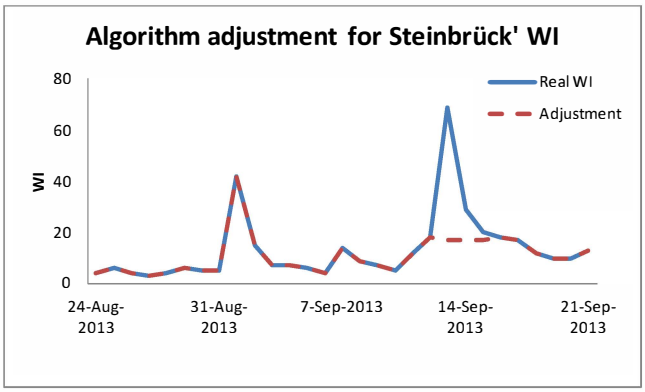

Fig. 5. Input data adjustment

D. Feedback from Historic Data

In order to further eliminate the noise, generated by factors such as each party's voters' profile, and in particular how familiar are a party's voters with the internet, a factor is calculated that connects the web search interest for a party and its electoral percentage. In cases where the search behavior of the electorate of each party does not change drastically between consecutive elections, in particular compared with the search behavior of the electorate of the other parties, then one can calculate this factor for the pre-election period of a previous election race and use it to predict the results of a forthcoming election race.

On the other hand in cases where major differentiations are observed in relation to the search behavior of the electorate of one party compared mainly with the relevant search behavior of the others parties, then the feedback from the previous election races could be ignored and the predictions could solely based on the behavior of the electorate during the pre-election period.

The calculation of the Web search Interest (WI) values in our dataset go through the following steps:

Let $WI_{N,\ party\ x,\ current\ elections}$, be the Web Interest value one day before the elections for Party x during the current elections race, and with N days duration for the observation window. First we calculate the Average Web Interest (AWI) over a

period of N days before the election date (in our scenarios N is equal to 30) :

$$AWI_{\text{party x, current elections, period}} = \frac{1}{N}\sum_{i=1}^{N} WI_{\text{i,party x,current elections,period}} \quad (1)$$

After calculating the AWI, we normalize the average web interests of each party to 100% percentage to arrive at the Normalized Web Interest (NWI) for each party, by calculating first the Total Web Interest (TWI):

$$TWI_{\text{current elections, period}} = AWI_{\text{party y, current elections, period}} + AWI_{\text{party x, current elections, period}} \quad (2),$$

$$NWI_{\text{party x, current elections, period}} = \frac{AWI_{\text{party x, current elections, period}}}{TWI_{\text{current elections, period}}} \quad (3)$$

Finally we divide the NWI of each party with the actual percentage of the party in the elections, normalized to 100%. The result is called Model Indicator (MI) and expresses the relation between the web interest for one party and the actual elections results of the party for a previous election race. The MI can be applied to the prediction of a future election race, if we choose to include historic data.

As described above in order to predict the normalized percentage (NP) of each party we follow two methods, one based on historic data and the other using only current data. The decision whether to use historic data is mainly based on the relation between the AWIs of the parties between the previous and the forthcoming elections.

In the case where historic data are used in order to predict the normalized percentage (NP) of each party in the next elections, we are using the MI of the previous elections and the AWI of the next elections.

$$NP_{\text{party x, current elections, period}} = NWI_{\text{party x, current elections, period}} * MI_{\text{party x, previous elections, period}} \quad (4)$$

In the case where historic data are ignored, the predictions are mainly based on the current values of the WI without feedback based on the MI of the previous elections. To do so we regard as the normalized percentage (NP) of each party in the next elections, the relevant NWI as it is calculated from current WI values.

In both methods, for each prediction of the normalized percentage, we compare the prediction of the model with the actual normalized percentage of the relevant election race after elections are held. Since we are not in a position to know or to predict with accuracy, before the elections, the actual sum of the percentages of the two major parties, we restrict model prediction to the normalized percentage for each party. Nevertheless, the prediction of the normalized percentage for each party represents among others, the percentage differences between the actual percentages of the parties and therefore is a strong indication for the actual results of an election race.

IV. PREDICTING THE ELECTIONS RESULTS

In Germany the three more recent national elections were held on September of 2005, 2009 and 2013.

	Elections 2005		***Elections 2009***		***Elections 2013***	
	CSU/CDU	*SPD*	*CSU/CDU*	*SPD*	*CSU/CDU*	*SPD*
WI	30,53	18,67	34,73	23,73	33,91	26,06
NWI	0,62	0,38	0,59	0,41	1,00	1,00
Results (actual)	35,20%	34,20%	33,80%	23,00%	41,50%	25,70%
Results (normilized)	50,72%	49,28%	59,51%	40,49%	61,76%	38,24%
Nor. predictions current wi	62,06%	37,94%	59,41%	40,59%	56,54%	43,46%
Error with current wi	-11,34%	11,34%	0,10%	-0,10%	5,21%	-5,21%
Nor. Predictions historic wi	na	na	47,94%	52,06%	56,65%	43,35%
Error with historic wi	na	na	-11,57%	11,57%	-5,11%	5,11%
Nor. Predictions mean wi	na	na	53,67%	46,33%	56,59%	43,41%
Error with mean wi	na	na	5,83%	5,83%	5,16%	5,16%

Table 2. Algorithm results

Table 2 contains for each election race the following information: the Web Interest, the Normalized WI for the selected set of terms, the election (actual and normalized) results, the algorithm's predictions and the prediction's error. It should be noted that the monitoring period for the results contained in Table 2 is 30 days. The algorithm provides three set of normalized predictions: one using historic data ("nor. Predictions historic wi"), one without taking into account the historic data ("nor. predictions current wi") and one with the average of the two above methods ("nor. predictions mean wi").

It is obvious that, in all cases of German national elections, the algorithm predicts the winner of the national election, and also predicts the normalized percentage of the two major parties with acceptable accuracy at least for the two more recent elections races.

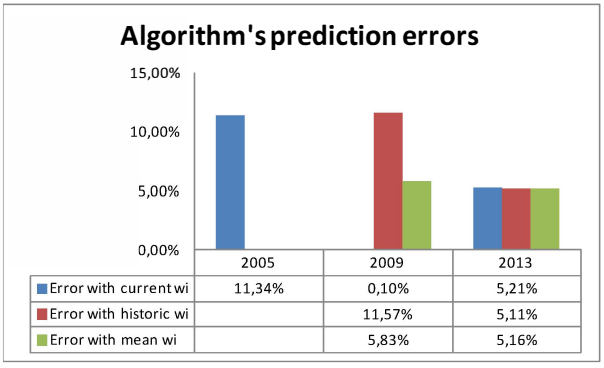

Fig. 6. Algortihm' s prediction errors

Following the rules described in section III in relation to the usage or not of historic data, the proposed algorithm predicts

the normalized percentages of the two major parties with accuracy 0,1% and 5,11% for 2009 and 2013 election races respectively. On the other hand although the algorithm predicts the winner of the 2005 election, it fails to predict with acceptable accuracy the parties' percentages. We believe that the high error of 2005 elections is due to low internet penetration of 2005 and consequently the low usage of web search engines.

V. CONCLUSIONS

The results of this work shows that the web-search engine behavior of potential voters during a short period before the elections can be connected with the final elections results. This correlation is so high as to allow the prediction of the results of forthcoming elections.

More specifically the paper shows that by applying a relevant algorithm to web search data, it is possible to predict the winner of forthcoming elections and to some extent the normalized percentages of the main parties. In particular a prediction algorithm was applied to three more recent national elections in Germany. In all cases the proposed algorithm predicts the winner of the elections by analyzing public web search statistics gathered from Google Trends. In addition the algorithm predicts with acceptable accuracy, the normalized percentages of the two major parties in each national election before the election date.

The results of our work indicate that web search based predictions may soon rival traditional polls despite the fact that the data released to the public by major search engines contain less demographic information compared to traditional polls.

Further research is required in order to examine the impact of all parameters in the results of the algorithm, in order to identify a set of parameters value that leads to results with the highest accuracy. In addition the algorithm should be tested and verified in more national elections aiming to determine all the crucial logical paths of the algorithm in order to eliminate noise and render the algorithm easily adaptable to any national election race.

REFERENCES

[1] Wired.co.uk, "Forget real time: 'next time' is far more disruptive", Wired.co.uk, 2012, DOI=http://www.wired.co.uk/news/archive/2012-02/20/forget-real-time, last accessed: March 2nd, 2012.

[2] S. Pion S. and L. Hamel L, "The Internet Democracy: A Predictive Model Based on Web Text Mining", In Proceedings of DMIN 2007, pp. 292—300

[3] S. Davidowitz, I. Seth, "Using Google Data to Predict Who Will Vote". Available at SSRN: http://ssrn.com/abstract=2238863

[4] X. Jin, A. Gallagher, L. Cao, J. Luo and J. Han, "The Wisdom of Social Multimedia: Using Flickr for Prediction and Forecast", In Proceedings of 2010 ACM Multimedia Int. Conference.

[5] S. Asur and A. B. Huberman. "Predicting the Future with Social Media", In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '10), Vol. 1. IEEE Computer Society. Washington, DC, USA, 492-499.

[6] D. Gayo-Avello, P. T. Metaxas and E. Mustafaraj E, "Limits of Electoral Predictions using Social Media Data", In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, July 17-21, 2011.

[7] S. E. Polykalas, G. N. Prezerakos, A. Konidaris, "A General Purpose Model for Future Prediction Based on Web Search Data: Predicting Greek and Spanish Elections", AINA Workshops 2013: 213-218