University of New Haven

Phase 3: Report a review on the related work

Team Report by Team OLAPPED CSCI 6401-01

Team Name :- OLAPPED

Course :- Data Mining

Instructor :- Prof. Shivanjali Khare

Course ID :- CSCI-6401-01

Session :- Fall 2023

Assignment :- Phase 3: Report a review on the related work

1. Team Name :- OLAPPED

Team Member Names :-

(1) Sean Vargas – svarg1@unh.newhaven.edu

(2) Kaylie N Neal – kneal5@unh.newhaven.edu

(3) Prashant Rana – prana4@unh.newhaven.edu

(4) Rajdeep Bhattacharya – rbhat6@unh.newhaven.edu

Parts of points from Phase 2: Selecting a research question & a dataset.

We are going to follow the following 8 steps to dodata mining for our project :-

1. Data Selection.

2. Data Cleaning.

3. Data Integration.

4. Data Reduction.

5. Data Transformation.

6. Data Mining.

7. Pattern Evaluation.

8. Knowledge Representation.

2. Selected Dataset :-

Description of the selected dataset that we want to work with –

The Data Set is named “trends” as it represents the Google Search Trends for a period of 20 years (2001 - 2020), pointing out the **[Top 5 Google Searches (Search Queries)]** by **[Categories]** with their **[Global Ranks]** and the ranks for the **[Top Countries]** by **[Year].**

It has 5 Attributes (represented by 5 columns) namely **location, year, category, rank and query**. And 26956 Data Points (represented by 26956 rows).

Research Question :-

The Research Question: - How can the patterns and trends in the **Google Search Engine Queries Dataset** be used to identify better rules for finding the most promising **Keywords (Search Queries)** and **Topics (Categories)** for showing more relevant **Search Results** in the future and targeting the relevant **SERPs (Search Engine Result Pages)** for **Ad Suggestions** by **timings (Seasonality)** and **locations (Geography)** for **Customer/Viewer Satisfaction** and higher **Ad Revenues**.

3. List of related reviews.

I. Title of the paper :-

A TAIEX forecasting model based on changes of keyword search volume on Google Trends.

The author(s) name and affiliation :-

Min-Hsuan Fan, Information Management, National Taichung University of Science and Technology, Taichung, Taiwan.

En-Chih Liao, Information Management, National Taichung University of Science and Technology, Taichung, Taiwan.

Mu-Yen Chen, Information Management, National Taichung University of Science and Technology, Taichung, Taiwan.

The publication date :-

Published in 2014 IEEE International Symposium on Independent Computing (ISIC)

Date of Conference: 09-12 December 2014.

The name of the publisher :-

IEEE.

APA Style Citation :-

Shibboleth authentication request. (n.d.). https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/7011756/authors

II. Title of the paper :-

Reinforcement Learning for Stock Price Trading with Keywords in Google Trends.

The author(s) name and affiliation :-

Shingchern D. You, National Taipei University of Technology, 1, Sec. 3, Chung-Hsiao East Rd, Taipei, Taiwan.

Po-Yuan Hsiao, National Taipei University of Technology, 1, Sec. 3, Chung-Hsiao East Rd, Taipei, Taiwan.

Shengzhe Tsai, National Taipei University of Technology, 1, Sec. 3, Chung-Hsiao East Rd, Taipei, Taiwan.

The publication date :-

Published in 2023 9th International Conference on Applied System Innovation (ICASI).

Date of Conference: 21-25 April 2023.

The name of the publisher :-

IEEE.

APA Style Citation :-

Shibboleth authentication request. (n.d.-a). https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/10179534/authors#authors

III. Title of the paper :-

An algorithm based on Google Trends' data for future prediction. Case study: German elections.

The author(s) name and affiliation :-

Spyros E. Polykalas, TEI of the Ionian Islands, Dept. of Business Administration.

George N. Prezerakos, TEI of Piraeus, Dept. of Electronic Computer Systems, Athens, Greece.

Agisilaos Konidaris, TEI of the Ionian Islands, Dept. of Business Administration.

The publication date :-

Published in IEEE International Symposium on Signal Processing and Information Technology.

Date of Conference: 12-15 December 2013.

The name of the publisher :-

IEEE.

APA Style Citation :-

Shibboleth authentication request. (n.d.-a). https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/6781856/authors#authors

IV. Title of the paper :-

Predicting Automotive Sales using Pre-Purchase Online Search Data.

The author(s) name and affiliation :-

Philipp Wachter, University of Hohenheim, Schwerzstr. 35, Stuttgart, Germany.

Tobias Widmer, University of Hohenheim, Schwerzstr. 35, Stuttgart, Germany.

Achim Klein, University of Hohenheim, Schwerzstr. 35, Stuttgart, Germany.

The publication date :-

Published in 2019 Federated Conference on Computer Science and Information Systems (FedCSIS).

Date of Conference: 12-15 December 2013.

The name of the publisher :-

IEEE.

APA Style Citation :-

Shibboleth authentication request. (n.d.-a). https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/8860027/authors#authors

V. Title of the paper :-

Recommending Personalized Search Terms for Assisting Exploratory Website Search.

The author(s) name and affiliation :-

Young Park, Department of CS&IS, Bradley University, Peoria, Illinois, U.S.A.

The publication date :-

Published in 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

Date of Conference: 02-06 June 2019..

The name of the publisher :-

IEEE.

APA Style Citation :-

Shibboleth authentication request. (n.d.-a). https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/8791155/authors#authors

4. Critical review of the related literature from different perspectives.

I. A TAIEX forecasting model based on changes of keyword search volume on Google Trends.

The selected data set :-

This study uses Google Trends as the research subject. The data provided by Google Trends on a weekly basis; the weekly statistics is from Sunday to Saturday.

The study was conducted in Taiwan and analyzes data on search volume provided by Google Trends.

In order to conform to the data provided by Google Trends, the research period is set to be from January 4, 2004 to June 29, 2013. The searched keywords were based on the randomly selected keywords from geographical names, finance-related topics, entertainments, names, and sports topic, with total of 103 keywords.

The number of sample records :-

The searched keywords were based on the randomly selected keywords from geographical names, finance-related topics, entertainments, names, and sports topic, with **total of 103 keywords**.

The data provided by Google Trends on a weekly basis; the weekly statistics is from Sunday to Saturday. Moreover, after comparing with TAIEX and excluding closing of Taiwan stock market during Chinese New Year, **each keyword has a value of 485 search volume**.

The TAIEX data included date, opening price, closing price, day high and day low. Through comparing with Google Trends search volume data, the **total number of TAIEX data is 485**; the study eventually just uses the closing price for analysis.

So the number of sample records is 103 X 485 = 49,955.

The research questions :-

How to analyze the change in keyword search volume in Google Trends and U.S. Dow Jones index to find a correlation between search behavior of local people and local stock market to master the stock market volatility and profit from it.

The data mining techniques :-

The data provided by Google Trends on a weekly basis from Sunday to Saturday. Moreover, after comparing with TAIEX and excluding closing of Taiwan stock market during Chinese

New Year, 103 randomly selected keywords, each keyword with a value of 485 search volume. The values of keyword search volume, after Google Trends formalization, distribute within the range of 0–100. TAIEX data collected is from January 4, 2004 to June 29, 2013.

The performance metrics :-

Pearson correlation coefficient.

The highest quantitative performance outcome :-

The experimental result of the TAIEX dataset shows the best return value of 1310.74.

II. Reinforcement Learning for Stock Price Trading with Keywords in Google Trends.

The selected data set :-

Google Trends Data - In August 2008, Google started a new service, called Insights for Search. This service allowed general users to track words or phrases typed in the Google's search box. Later, in 2012, the Insights for Search was merged to Google Trends.

As the Google Trends provides the relative counts of a word (or phrase) appeared in the search box, it is correlated to public interests or concerns. Therefore, it is a suitable tool to probe the general interests in financial events.

The number of sample records :-

A. Experiment One :-

During training and testing, the agent trades stocks or ETFs once per week (five trading days). In addition, the time series data (i.e., interest over time) from Google Trends are collected in a yearly basis.

In this experiment, target stocks or ETFs are used in trading. The collected pricing information is from 05/2017 to 04/2022 from Yahoo Finance.

The first three years are used for training, the fourth (or fifth) year is used for validation, and the rest one year is for testing. During training, 100 agents are trained. Ten out of the 100 best agents are selected based on the performance on the validation year. These agents are used for testing and the average performance of these agents is reported as the experimental result.
B. Experiment Two :-

In this experiment, a downloaded (normalization) period consisted of two months of trading period plus previous 0.5 month and next 0.5 month. For example, if the trading days are from March 1 to April 30, the downloaded observation period is from Feb. 15 to May 15. Then it is applied for time-series data from March 1 to April 30.

The research questions :-

There are two experiments that have been mentioned in this research paper.

Experiment One was intended to study if the agent can take advantage of using the strengths of the keywords to obtain better trading performance.

Experiment Two used the the ticker symbol of a stock as the keyword to study the benefits of adding the keyword strength during trading.

The data mining techniques :-

The first three years are used for training, the fourth (or fifth) year is used for validation, and the rest one year is for testing. During training, 100 agents are trained. Ten out of the 100 best agents are selected based on the performance on the validation year. These agents are used for testing and the average performance of these agents is reported as the experimental result.

The performance metrics :-

Experiment One :-

Keywords correlations with the trading performance.

Experiment Two :-

The underlining concept of this approach is to use the keyword strength as a measure of the public attention.

Then they tried to find the correlation (if any) between public attention (i.e. keyword strength) and the return of the proposed approach.

The highest quantitative performance outcome :-

AAPL receives a lot of attention, followed by QCOM.

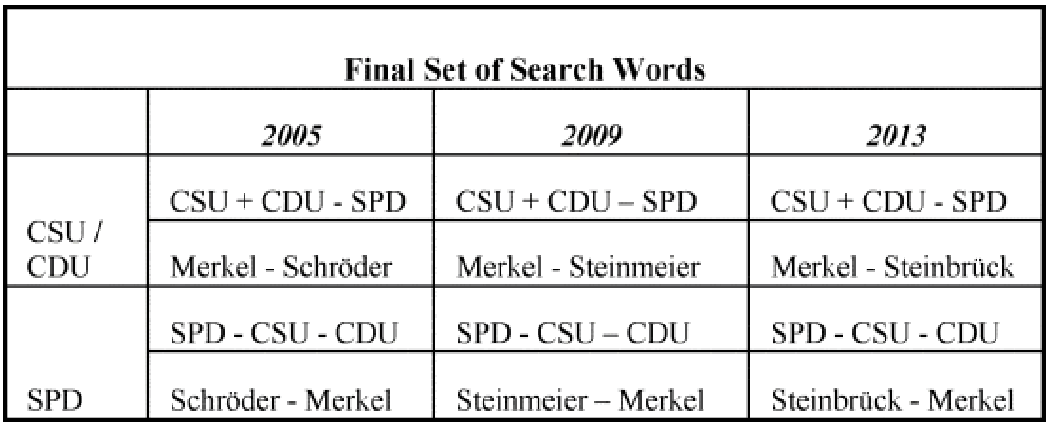
III. An algorithm based on Google Trends' data for future prediction. Case study: German elections.

The selected data set :-

The analysis of information that is provided by popular search engines (Google, Yahoo etc) with respect to the volume of searches for specific terms.

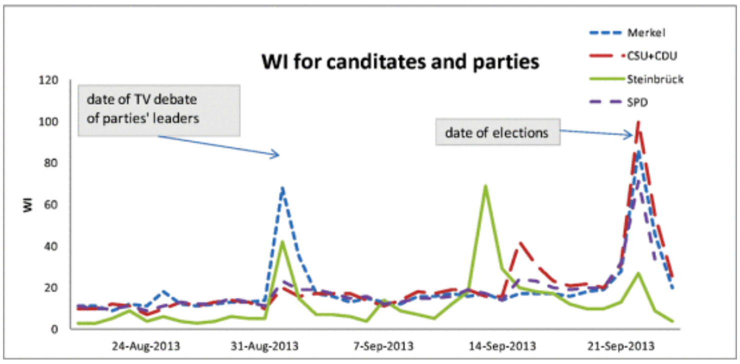
In this paper an algorithm has been applied to the data provided by the Google Search engine via the **Google Trends service**.

The number of sample records :-

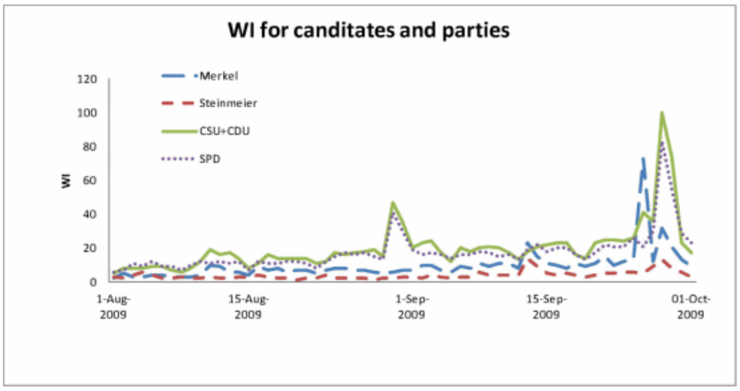


Final Set of Search Words			
	2005	*2009*	*2013*
CSU / CDU	CSU + CDU - SPD	CSU + CDU – SPD	CSU + CDU - SPD
	Merkel - Schröder	Merkel - Steinmeier	Merkel - Steinbrück
SPD	SPD - CSU - CDU	SPD - CSU – CDU	SPD - CSU - CDU
	Schröder - Merkel	Steinmeier – Merkel	Steinbrück - Merkel

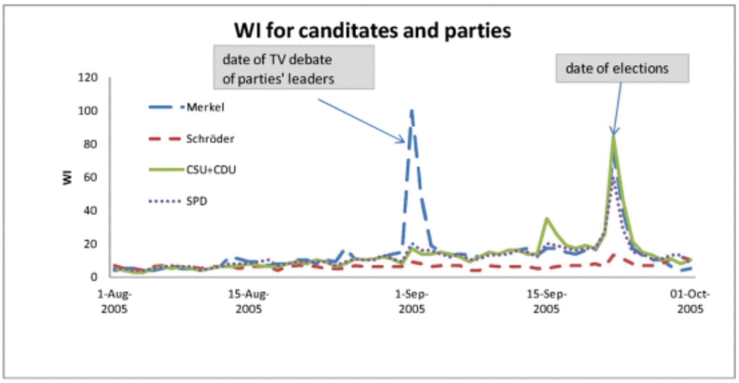
The above table contains the final set of search words for each election race.



The figure above depicts the Web Interest values of the selected terms for the 2013 elections.



The figure above depicts the Web Interest values of the selected terms for the 2009 elections.



The figure above depicts the Web Interest values of the selected terms for the 2005 elections.

The research questions :-

This paper examines the relation between the search preferences of web users and the results of the German national elections of 2005, 2009 and 2013. The analysis is focused on the selections of the appropriate set of search terms, in an appropriate timeframe to arrive at an accurate estimate of the election results with respect to the two major parties in Germany.

The data mining techniques :-

Data selection.

Data cleaning.

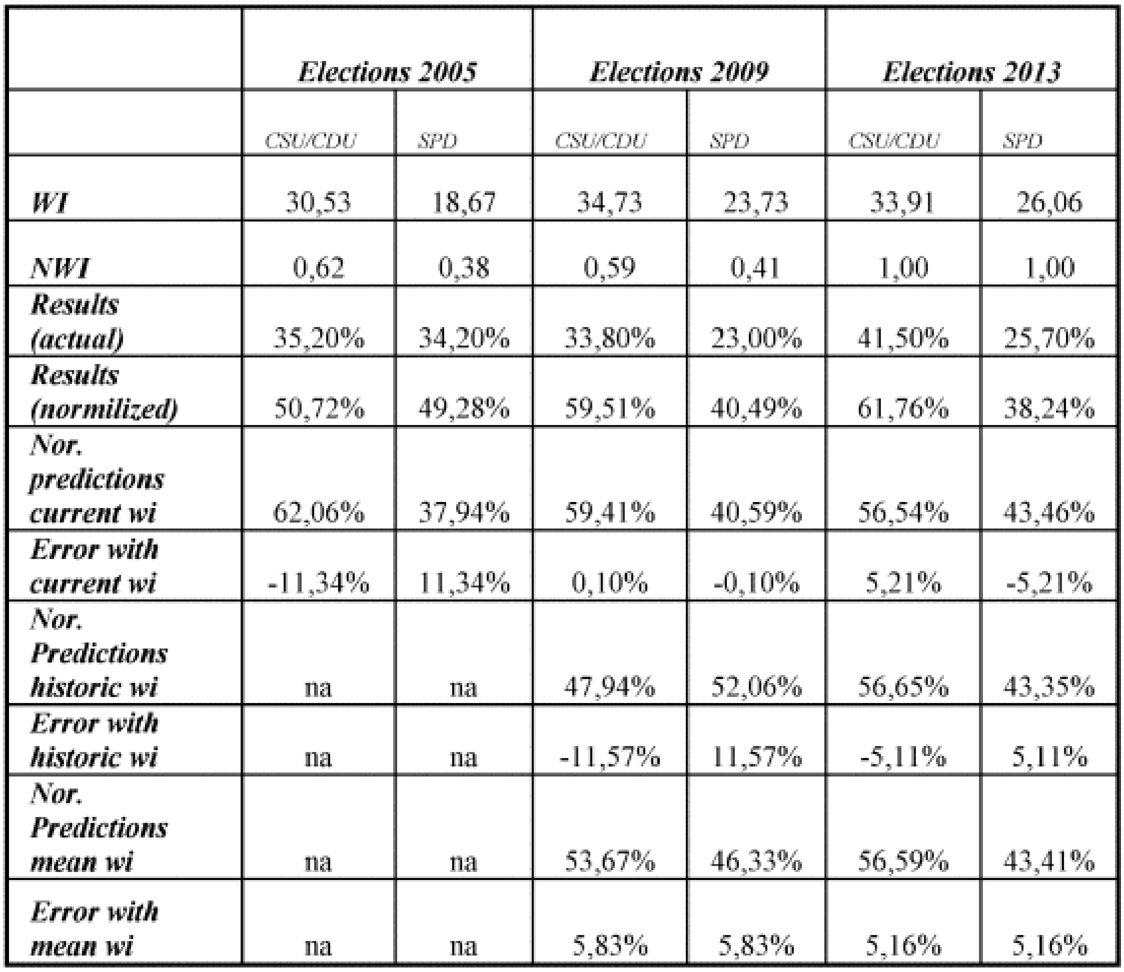
Normalization.

Time frame analysis.

The performance metrics :-

Normalized predictions and the prediction's error with taking into account historic data and without taking into account historic data.

The highest quantitative performance outcome :-



	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%

Results.

IV. Predicting Automotive Sales using Pre-Purchase Online Search Data.

The selected data set :-

In 2006, Google launched the search analysis website Google Trends. The publicly available tool provides information about aggregated individual searches expressed in a search volume index. Hence, Google does not report the data in absolute numbers but provides the relative popularity of a search term. The index is calculated by dividing the data points of a query by the total volume of searches of the geography and time range considered. The query shares are normalized, such that 100 indicates the highest query share of whole period. It allows to compare the relative popularity of a query across different geographic locations and time intervals. Google also introduced different categories and subcategories to refine the search for terms with multiple meanings. In the context of the automotive industry, the search results for “beetle” can be narrowed down by the choice of an appropriate category to exclude queries regarding the insect and only obtain results for the car of by Volkswagen.

The number of sample records :-

We collected monthly search query indices for the respective car model and/or car brand in combination with the most relevant keywords selected via Google Ads. We focus on the car manufacturer Honda as a representative of a large seller in the US. To obtain Google Trends data for the brand Honda, we additionally include the model names of the four best-selling car models responsible for approximately 90% of the Honda car sales in the period considered.

The result data are limited to searches originating from the US in the period from January 2004 to February 2019.

The research questions :-

This research paper describes the experimental evaluation of a forecasting technique for car sales based on most relevant Google Trends data.

The data mining techniques :-

Data selection.

Data cleaning.

Normalization.

Time frame analysis.

Adjustments for adding seasonality.

The performance metrics :-

Correlation coefficients, Root mean squared error and Mean absolute error.

The highest quantitative performance outcome :-

The strongest correlation with car sales was observed for S & P 500 without time lag.

V. Recommending Personalized Search Terms for Assisting Exploratory Website Search.

The selected data set :-

Not mentioned.

The number of sample records :-

Not mentioned.

The research questions :-

This paper describes a personalized word recommender system called mySearchCLUE for potential search terms within a website search engine using collaborative filtering, in which the users are visitors, and the items are words and webpages.

The data mining techniques :-

Data selection.

Data cleaning.

Normalization.

Time frame analysis.

The performance metrics :-

Pearson correlation coefficient.

The highest quantitative performance outcome :-

Not applicable.