CHARLES UNIVERSITY IN PRAGUE

FACULTY OF SOCIAL SCIENCES

Institute of economic studies

Kristýna Brunová

Are the more popular stocks also the more risky ones?

Google and Wikipedia searches in portfolio optimization

Bachelor Thesis

Prague 2015

Author: **Kristýna Brunová**
Supervisor: **PhDr. Ladislav Krištoufek, Ph.D.**
Academic year: 2014/2015

Bibliografický záznam

BRUNOVÁ, Kristýna. *Are the more popular stocks also the more risky ones? Google and Wikipedia searches in portfolio optimization.* Praha 2015. 45 s. Bakalářská práce (Bc.) Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií. Vedoucí bakalářské práce: PhDr. Ladislav Krištoufek, Ph.D.

Anotace (abstrakt)

Tato práce zkoumá, zda je možné použít data z Google Trends a Wikipedie k diverzifikaci portfolia. Je založena na empirickém faktu, že nárůst v objemu vyhledávání souvisejícím s konkrétní akcií je spjat s vyšší volatilitou cen této akcie. Použijeme tedy diverzifikační strategii, která diskriminuje populární akcie tím, že jim přiřazuje nižší váhy v portfoliu. Na druhou stranu, nejméně hledané akcie jsou v portfoliu preferovány. Abychom zjistili popularitu akcie, zaměříme se na objemy vyhledávání na Google pro slova související s akcií, a také na to, kolikrát si lidé prohlédli stránku odpovídající dané společnosti na Wikipedii. Výsledky studie ukazují, že strategie založené na objemech vyhledávání dosahují nižšího rizika a vyšších standardizovaných průměrných výnosů než srovnávací index a portfolio, kde mají všechny akcie stejnou váhu. Tyto strategie jsou navíc úspěšné i out-of-sample.

Klíčová slova

Diverzifikace portfolia, Google, Wikipedia, vyhledávání na internetu, online pozornost

Abstract

This thesis studies if the web search data provided by Google Trends and Wikipedia can be utilized for portfolio diversification. We build up on the empirical results indicating that the surge in online attention paid towards a specific stock is associated with an increase in the stock price volatility. Therefore, we employ a diversification strategy that discriminates for the popularity of a stock by assigning it a lower portfolio weight. Conversely, the least searched stocks are preferred in the portfolio. To measure the popularity of a stock, we focus on Google search volume for stock-related terms as well as on Wikipedia pageviews of the corresponding company's page. Our results show that the search-based strategies outperform the benchmark index and the uniformly distributed portfolio, reaching lower risk level and higher standardized average returns. Moreover, these strategies are successful even in the out-of-sample.

Keywords

Portfolio diversification, Google, Wikipedia, internet search, online attention

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 15, 2015

Signature

Acknowledgments

I am heartily thankful to my supervisor, PhDr. Ladislav Krištoufek, Ph.D., for his patience, guidance and valuable comments.

Contents

1 Introduction

In the last decade researches have obtained access to a huge amount of information about what people search on the Internet through interfaces like Google Trends, Baidu search or Wikipedia Trends. These services created a perfect gateway to achieve a wholly new perspective on the real-word phenomena, since the search data showed up to be useful in explaining multiple aspects of people's behavior. Accordingly, growing attention has been given in academic literature to these data. The application ranges from influenza tracking (Polgreen et al., 2008; Ginsberg et al., 2009) and unemployment predictions (Choi and Varian, 2009b; D'Amuri and Marcucci, 2010) to the search data utilization in finance (Preis et al., 2010; Bank et al., 2011; Vlastakis and Markellos, 2012).

In this paper we examine whether the web search data provided by Google and Wikipedia can be useful for portfolio diversification. Diversifying portfolio is essential for eliminating the security-specific risk, thus reduces portfolio loss and volatility, and maximizes risk-adjusted return. We apply a diversification strategy introduced by Kristoufek (2013b) to Google and Wikipedia stock-related searches and subsequently examine the performance of the search-based portfolios. The strategy stems in an idea that the more searched stocks are also the more volatile, hence exhibit higher risk. Such stocks should be therefore discriminated in the portfolio by assigning them lower weights. On the other hand, preferring less popular (or peripheral) stocks is expected to result in lower variance portfolios. A sample of twenty-two Dow Jones Industrial Average constituents creates the initial dataset.

To measure the stock's popularity on the Internet, we focus on these searched keywords: first, we follow Da et al. (2009) and Kristoufek (2013b) and utilize Google search volume for ticker symbols. Next, we take the ticker symbol and combine it with the word stock (e.g. "stock XOM"). Finally, we employ data on pageviews of the English-language Wikipedia for articles dedicated to the DJIA companies.

Most authors used search volume data of a weekly frequency (Kristoufek, 2013b; Preis et al., 2010; Vlastakis and Markellos, 2012). Here, we focus on daily data for the following reasons: first, we study whether the results of Kristoufek (2013b) obtained for the Google Trends based strategies apply also at a daily timescale. Second, we are aware of losing some information if weekly data are employed.

The purpose of this thesis is to investigate whether the portfolios constructed in line with the proposed diversification strategy are able to reach lower risk level (measured by standard deviation) and higher standardized returns (measured by the Sharpe ratio). The performance of the search-based portfolios is compared with the uniformly weighted portfolio and the benchmark DJIA index.

The remainder of this thesis is organized as follows. Section 2 provides a review of the most important related papers. Section 3 describes the data and sample construction. Section 4 explains the diversification strategy and introduces all scenarios established for the construction of portfolios. Section 5 presents and discusses the results. Finally, Section 6 concludes.

2 Literature review

This section is divided into two parts. The first part summarizes influential applications of web search data in various scientific fields. The second part focuses on studies concerning the utility of Internet-related data in finance.

2.1 Non-financial application of web search data

To the best of my knowledge, the first published paper that discovered the utility of web search data was Ettredge et al. (2005). The authors found a link between the search frequency of job-related terms and the official U.S. monthly unemployment data, suggesting a potential predictive power of web-based data for important macroeconomic statistics. At about the same time Cooper et al. (2005) suggested Internet search terms related to specific cancers to be associated with the estimated incidence and mortality for these cancers. Since then many authors tried to examine the relationship between web search data and various indicators from distinct fields. For example, in the field of epidemiology, Polgreen et al. (2008) and Ginsberg et al. (2009) in one of the most successful applications in this direction showed that influenza-related web search traffic can be used to accurately track influenza spreading as measured by official data on contagion collected by Health Care Agencies. Their work was widely publicized and stimulated several other papers concerning the utility of web search data in estimating flu activity (Carneiro and Mylonakis, 2009; Hulth et al., 2009; Corley et al., 2009; Dugas et al., 2012). These findings also led to the inception of Google Flu Trends[1], a freely-available service that helps to predict influenza outbreaks. Recently, Preis and Moat (2014) further demonstrated the supremacy of Google Flu Trends data in real-time influenza monitoring over the official reports of flu infections, even when these reports are available with only one week's delay. Web search data application in epidemiology was also extended to other illnesses, most notably by Brownstein et al. (2009), Pelat et al. (2009), Wilson (2009) and Zhou et al. (2011).

In economics, research related to web search data essentially started with the public release of the Google Trends portal in 2008. Shortly after this public release, Google chief economist Hal Varian along with Hyunyoung Choi published a paper suggesting Google Search Insights data[2] utility in

[1] Available via https://www.google.org/flutrends/

[2] There were two user interfaces for the data, Google Trends and Google Insights for Search, which merged into one in 2012. Choi and Varian (2009) used Google Insights for

predicting the present which is, as authors noted, a form of 'contemporaneous forecasting' or 'nowcasting' (Choi and Varian, 2009a, b). The authors used several examples to demonstrate the predictability of Google Trends data including initial claims for unemployment benefits, home and automotive sales and vacation destinations.

Besides Choi and Varian (2009b), many authors examined web search data's ability to predict unemployment, including D'Amuri and Marcucci (2010) in the US, Suhoy (2009) in Israel, Askitas and Zimmermann (2009) in Germany and McLaren and Shanbhogue (2011) in the UK.

Recently, Pavlicek and Kristoufek (2015) extended the application on unemployment rates of smaller countries, specifically those forming the Visegrad Group (i.e. the Czech Republic, Hungary, Poland and Slovakia). The results they obtained are quite diverse: adding job-related Google searches into the model for nowcasting the unemployment rate strongly enhances the model in case of the Czech Republic and Hungary. For Poland, a statistically significant difference of forecasting accuracy between the "Google model" and the base model (i.e. the model without the incorporation of the Google series) is present only for the maximum lag of 6 months, and for Slovakia, the base model even outperforms the "Google model". Nevertheless, the authors demonstrated that Google searches can be utilized in nowcasting the unemployment rates even in the countries distinct from the well-developed (Western) ones.

Baker and Fradkin (2011) focused on the elasticity of web search activity with respect to unemployment benefits. Their results indicate that the increase in duration of unemployment benefits leads to negative, but small job search responses in most specifications. Furthermore, they found that moderate magnitudes and short duration of the job search responses suggest that persistent unemployment in the Great Recession was not primarily driven by the extensions of unemployment benefits. Another macroeconomic application, suggested by Guzman (2011), utilized Google search data for the inflation prediction.

Lastly, Della Penna and Huang (2009), Carrière-Swallow and Labbé (2011) and Vosen and Schmidt (2011) utilized search results for retail goods to nowcast private consumption; the result being that their constructed indices outperform the survey-based indices currently in use. For example,

Search, because it allowed a logged-in user to download search data as a CSV file. From now on, the term Google Trends is used for either Google Insights for Search or Google Trends database.

Della Penna and Huang (2009) created consumer sentiment index for the U.S. based on the popularity trends of selected Google searches, and showed that such an index can be a useful leading indicator in forecasting consumer spending among the non-search-based indices used by authors for comparison.

2.2 Financial application of web search data

Since this thesis examines the utility of web search data for financial market participants, the following section tries to summarize important papers dealing with stock market related application of web search data.

Several papers studied the relationship between web searches and market activity with the aim to confirm the hypothesis that volume shifts of market-related searches can affect or even anticipate fluctuations in financial market indicators. To my knowledge, Preis et al. (2010) were the first to examine the link between Google search volume and financial market fluctuations. By taking weekly transaction volumes of S&P 500 companies, they found a clear evidence that these transaction volumes are cross-correlated with weekly search volumes of corresponding company names. The authors also applied a method introduced in Preis et al. (2008) for quantifying complex correlations in time series, discovering that search volume time series and transaction volume time series show recurring patterns, indicating that there is a clear link between weekly transaction volumes and weekly search volumes. Moreover, as authors pointed out, *"there is not only a linear dependence but also complex dependencies, which raises hopes that search volume data can contribute to understand financial crises"* (Preis et al., 2010, p. 5718).

Bordino et al. (2012) confirmed the results of Preis et al. (2010) also at a daily time scale, although they used Yahoo!, not Google, daily data to investigate whether today's query volumes can anticipate tomorrow's financial indicators such as trading volumes, daily returns, volatility, etc., and found a significant anticipation for trading volumes. They do so by means not only of a time-lagged cross-correlation analysis, but also by the Granger-causality test. Their results suggest that web search data are able to Granger-cause trading volumes, but the opposite direction relationship is much weaker.

Dimpfl and Jank (2012) contribute to the literature by examining Google Trends data ability to predict stock market volatility. Their findings show that search queries Granger-cause volatility as today's increase in search queries for the keyword “dow” is associated with the surge in Dow Jones' Industrial Average volatility tomorrow. To utilize their results, the authors incorporated information from search queries into several prediction models for realized volatility and get more precise in- and out-of-sample forecasts, particularly in the long run and in high volatility periods.

Vlastakis and Markellos (2012) studied how idiosyncratic (i.e., firm specific) and market-related information demand and supply affect stock market activity and risk aversion. To proxy for the information demand, they took weekly search volume time series from Google Trends database using firm-specific keyword (e.g., “procter gamble” for Procter&Gamble Co., “disney” for Walt Disney Co.) for idiosyncratic information demand and “S&P 500” as the search keyword for market-related information demand; information supply was proxied by the financial information in the news. Their results show that market information demand is positively related to historical and implied measures of volatility and to trading volume. Moreover, they provide evidence for a significant effect of variations in information demand on historical volatility and trading volume at both the individual stocks and overall market level; however, the effect was stronger for market-related information demand. Their analysis also suggests that information demand tends to escalate during high-return periods and along with the investors' level of risk aversion.

Alanyali et al. (2013) examined the relationship between financial news and the stock market from a slightly different perspective; instead of using web search traffic as a proxy for the interest in a company, they analyzed a corpus of daily issues of the Financial Times, uncovering that a greater number of mentions of the company in the news corresponds to a greater transaction volume of company's stocks. They also found a link between the daily number of mentions of a company's name and the daily absolute return of the corresponding company's stocks; more precisely, the daily number of mentions of a company's name is positively related to daily absolute return of the company's stocks. However, their analysis provides no evidence that interest in a company in the news is correlated with company's stock price movements when direction of movement is considered.

The authors of Preis et al. (2013) questioned Google Trends data ability to anticipate certain future trends and obtained support for the hypothesis that notable drops in financial markets are preceded by periods of investor concern and, accordingly, Google search volumes for finance-related terms could have been exploited in the construction of profitable trading strategies. Their approach was to utilize search volume data in the construction of investment strategies, which stem in an idea that one should sell the DJIA at the closing price on the first trading day of week t and buy the DJIA at the end of the first trading day of the following week if the interest in a specific keyword, as measured by Google Trends, increased in the weeks preceding the week t. Logically, if the relative change in search volume measured in week $t - 1$ is negative, the opposite action - first buy then sell - should be done[3]. With respect to their results, the best performing Google Trends strategy was the one based on the search term debt, and overall, Google Trends strategies yielded significantly higher returns than the random strategies, i.e., the strategies based on buying and selling the market index in an uncorrelated, random manner. Moat et al. (2013) expanded their analysis to Wikipedia usage patterns by demonstrating that changes in Wikipedia usage patterns can be linked to subsequent stock market moves. In line with the findings of Preis et al. (2013), their results suggest that an increase in information gathering is followed by a decrease in stock prices. The authors implement a hypothetical investment strategy introduced in Preis et al. (2013), although they utilized data on Wikipedia page views and page edits for articles concerning DJIA companies and for those related to more general financial topics. In both cases, returns of Wikipedia page view based strategies were significantly higher than returns of random strategies, however, no significance difference between the returns of Wikipedia edit based strategies and the random strategies was obtained. Furthermore, following a decrease in Wikipedia page views related to financial topics, they found a mean DJIA weekly return significantly greater than zero, and, in contrast, following a surge in these views, they found a mean DJIA weekly return significantly lower than zero. The assumption that only the changes in Wikipedia page views of articles with financial connotations can provide insights into the information gathering processes which precede trading decisions, was further verified by a parallel analysis that found no relation between changes in views of Wikipedia articles related to actors and filmmakers and changes in the DJIA.

Curme et al. (2014) built on the results of Preis et al. (2013) and Moat et al. (2013) and provided further insights into the early information gathering stages of decision making processes. They focused on identifying topics that people search before stock market moves. By using data from Google and Wikipedia as well as judgments from the users of the online service Amazon

[3] The authors called such a strategy the Google Trends strategy.

Mechanical Turk, they found evidence that an increase in web searches relating to business or politics tends to precede stock market falls, although the predictive power of these search terms has recently diminished, indicating increasing employment of web-based data in automated trading strategies.

A lot of studies posed a question whose attention Google Trends, particularly its Search Volume Index, captures. Most of these studies showed that Google Trends likely captures the attention of individual investors and with this being established, they provided support to the Price Pressure Hypothesis of Barber and Odean (2008). This hypothesis assumes that individual investors are net buyers of attention-grabbing stocks, implying that an increase in individual investors' attention leads to a temporary positive price pressure. Their reasoning goes as follows: there are thousands of stocks which individual investors can potentially buy, but only a few they can sell, given that they can only sell stocks they already own, and therefore, to overcome this searching problem when buying stocks, individual investors tend to buy attention-grabbing stocks, e.g., stocks in the news, stocks experiencing high abnormal trading volume and stocks with extreme one-day returns. In contrast to individual investors' buying behavior, institutional investors are not as influenced by attention, because they routinely sell short (contrary to individual/retail investors), which makes the alternatives they can choose from when selling equal to ones they search for when buying. Moreover, institutional investors devote more time to search than do most of the retail investors, and they have access to more sophisticated information services; plus they usually own many more stocks than individuals do, thus increasing their selling set.

Barber and Odean's contribution was meaningful, although they used indirect proxies, e.g., stocks experiencing extreme returns or abnormal trading volume, to recognize stocks that more likely capture investor's attention. To my knowledge, Da et al. (2011) were the first to employ a direct measure of investor attention, when they proxied attention by the Google's Search Volume Index (GSVI). The authors argued that the GSVI likely captures attention of less sophisticated individual or retail investors and, using a sample of Russell 3000 stocks, they also found a positive, but small correlations between the GSVI and other proxies for attention, such as extreme returns, turnover and news. By focusing on abnormal Search Volume Index (ASVI)[4] as a measure of retail investors' attention, Da et al. (2011) provide empir-

[4] Da et al. (2011) defined ASVI as the (log) SVI during the current week minus the (log) median SVI during the previous eight weeks.

ical support to the attention-induced Price Pressure Hypothesis of Barber and Odean (2008), as an increase in weekly ASVI is associated with positive price pressure in the next 2 weeks which is, however, almost completely reversed by the end of the year. Joseph et al. (2011) obtained similar results for a different sample of firms (S&P 500), specifically, the current surge in search volume forecasts abnormal returns and increased trading volumes in the subsequent period. Moreover, they found the sensitivity of returns to the search volume to be lowest for easy-to-arbitrage, low volatility stocks and highest for difficult-to-arbitrage, high volatility stocks.

Contrary to Da et al. (2011), who focused on the search volume of a firm's formal ticker symbol, Bank et al. (2011) employed a more general approach by using weekly or monthly search volume of ordinary firm names as a proxy for investor attention because, arguably, the average Internet user will search for a company name rather than for a company's ticker symbol and, as the authors pointed out, this approach captures the information attention from *"a much broader and potentially relevant audience"* (Bank et al., 2011, p. 240). Accordingly, on a sample of German stocks, they demonstrated that an increase in search queries is associated with higher trading activity, improved stock liquidity and higher future returns in the short-run. The authors attributed this temporary interdependence between changes in search volume and stock return to the price pressure driven by individual investors, which is in accord with Barber and Odean (2008) and with empirical findings of Da et al. (2011). Furthermore, they reason the improvement in liquidity, with respect to market microstructure theory, as a consequence of the reduction in asymmetric information costs, and therefore conclude that search volume primarily measures the interest of uninformed investors, which is also in line with Da et al. (2011).

Shi et al. (2012) empirically verified the Price Pressure Hypothesis of Barber and Odean (2008), showing that the increase in individual investors' attention over a stock significantly elevates its turnover in the short-run; therefore produces upward pressure on the stock price. Moreover, they proved that the investment return of noise traders declines as the attention rises, which is also in accord with theoretical conclusions of Barber and Odean (2008). In performing the analysis, they used Baidu search data as well as analysts' neutral ratings, which Shi et al. employed to separate the effect of attention on stock price from direct effects of analysts' recommendations as much as possible.

Smith (2012) questioned whether changes in Google search volume for particular keywords can predict volatility in the foreign currency market. His findings confirmed the hypothesis that evolution in Google searches, at least for some keywords, significantly affects volatility, as he found a link between searches for the keyword “economic crisis+financial crisis” and week-ahead volatility for all seven currencies and also between the keyword “recession” and week-ahead volatility for five currencies.

Ramos et al. (2013) contribute to the research in investor attention by investigating the ability of web search queries measured by GSVI[5] to predict market behavior of 50 largest European stocks. They found that an increase in web search queries precedes a surge in liquidity and volatility, and a drop in cumulative returns; however, the effect on liquidity and volatility is only short-lived and is sharply reversed in the following week, indicating the presence of less sophisticated investors. Moreover, the predictive power of web search queries for returns and liquidity is intensified, when either the firm or the market hits a 52-week high, and mitigated, when the market hits a 52-week low. As for the limited attention theory of Peng and Xiong (2005), Ramos et al. (2013) provide empirical evidence that investors are more prone to process market information than firm specific information in investment decisions, suggesting category-learning behavior of these investors which, according to Peng and Xiong (2005), stems from the limited amount of attention they can devote to each investment decision.

The attention theory of Barber and Odean (2008) was also supported by findings of ap Gwilym et al. (2013), who proxied Chinese investors’ speculative demand by search interest in the word “concept stocks”. The authors used Chinese stock indices and found a positive, but contemporaneous, relationship between speculative demand and returns for these indices, which, as authors argued, is due to the price pressure driving up current price and returns. However, the rise of current price is only short-lived, and when the mispricing is corrected as the price falls, future returns decline. The same enhancement, stemming from the increased speculative demand, was obtained for trading volume. Furthermore, they provide additional support to the hypothesis that Google Trends measures the interest of retail, rather than institutional investors, as they found the effect of searches on returns and trading volume to be stronger for constituents of Chinese A shares indices (dominated by retail investors), than for those of Chinese B shares indices (dominated by institutional investors).

Zhang et al. (2013) quantified investor attention using search frequency

[5] Google Search Volume Index

of stock names in Baidu Index. By employing data from different boards of Chinese Stock Market, the authors showed that *"investor attention is a desired variable to predict stock abnormal return"* (Zhang et al., 2013, p. 617), because of the contemporaneous correlation between these two variables. Moreover, they demonstrated that investor attention Granger-causes abnormal return, while the opposite direction is relatively less obvious, and explained the Granger-causality from investor attention to abnormal return in terms of overconfidence; that is, the investor puts more trust in his or her private information and underestimates the accuracy of the public information, which may cause the public information not being reflected in price movements.

An interesting point to the discussion was brought by Mondria et al. (2010), who used attention allocation in explaining home equity bias. The authors were the first to measure attention allocated to a specific country by the number of times this country provides the answer to a search query, and showed that agents tend to increase their holdings of a particular country's assets if they have more information about that country, i.e., if more attention is allocated to the country.

Kristoufek (2013a) used Google Trends data as well as data on Wikipedia page views to proxy for investors' sentiment, which is a core variable for explaining the BitCoin currency price movements. The results he obtained show that there is an outstanding positive correlation between the BitCoin prices and both searches for the word BitCoin on Google and daily Wikipedia views of the page 'BitCoin'. He also uncovered that the nature of this relationship is bidirectional, concluding that *"speculation and trend chasing evidently dominate the BitCoin price dynamics"* (Kristoufek, 2013a, p. 5). Finally, to distinguish between investors' interest coming from positive events from the attention reflecting bad events, Kristoufek defined a positive feedback if the investors' interest (measured by search queries) is increasing in response to the above-the-trend-level price of BitCoin and a negative feedback if the surge in attention follows the price decline below its trend value. The subsequent analysis reveals that if the prices are growing and the corresponding public interest is increasing, prices will likely grow further. Conversely, if the prices are falling, the increased interest pushes them even deeper.

More recently, Vakrman and Kristoufek (2015) questioned Google Trends data usefulness in explaining two stylized facts of the initial public offerings (IPOs) - the long-term underperformance and the high initial returns (IPO underpricing). Using the sample of 75 initial public offerings that occurred in the USA between 2004 and 2010, they confirmed the presence of higher initial returns for IPOs receiving above-average attention (measured by the firm-specific Google search volume). This effect was, however, significant only for IPOs which took place in the periods of positive sentiment. Nevertheless, they showed that these high-attention IPOs, which were realized in the positive sentiment periods and yielded superior returns, experience a long-term price reversal. Furthermore, they provided support to the hypothesis that Google search volume is able to predict only one part of the initial returns - market overreaction to the offering -, while the other - the true IPO discount - does not seem to be influenced by attention in any direction.

Finally, the most influential paper for my thesis was Kristoufek (2013b), who utilized Google search data for portfolio diversification, which stems in an idea that the popularity of a stock is positively related to its riskiness through increased volatility, and therefore such popular stocks should be discriminated by assigning them lower weights in the final portfolio. The least popular stocks, on the other hand, should be assigned the highest weights. The popularity of a stock in a particular week is measured by weekly search volume of a stock-related term; for these terms Kristoufek used the firm's formal ticker symbol as well as the combination of the word "stock" and ticker symbol. The results he obtained show that the search-based strategies outperform the passive buy-and-hold strategy in terms of the Sharpe ratio and they are able to reach lower risk level than the uniformly weighted portfolio; however, better portfolio performances were achieved for the ticker strategy which is, as Kristoufek pointed out, due to investors are probably searching only for the ticker symbol rather than combining it with the word "stock".

3 Data and sample construction

This section is divided into two parts. The first part describes the nature of search volume data used throughout the paper. In the second part the sample construction is discussed.

3.1 Search volume data

Google Trends data

In May 2006 Google launched Google Trends, a tool that enables its users to see how popular certain search terms are across various geographic regions, cities and languages[6]. Shortly after this public release[7], Google came with the more sophisticated service called Google Insights for Search. Insights for Search, as opposed to Trends, allowed the logged-in users to slice the data by categories to distinguish between, e.g., searches for "apple" the fruit and "Apple" the company. Users could also assess the popularity of miscellaneous queries people type into Google's search box across finer geographic areas than with Trends and view it on a map. And arguably the biggest advantage of Insights over Trends was its possibility to download the data onto spreadsheets and use them in forecasting or research[8]. In September 2012, these two versions merged into one named Google Trends, which includes features from both Insights and previous Trends.

Now Google Trends[9] is a keyword research tool that provides information on how often a particular search term was entered into Google in a chosen world region, time period and Google Trends category[10]. One can also query Google Trends for searches including disjunctions of terms, related searches or those containing one search term and exclude the other: "Jobs - Steve" will return search intensity for searches containing the word "Jobs" while excluding those with the word "Steve".

Data on search frequency ranges back to the beginning of 2004 and are often referred to as Google Search Volume (GSV)[11]. The data, however, does not show the absolute search volume for a keyword, since Google Trends

[6] http://googleblog.blogspot.cz/2008/06/new-flavor-of-google-trends.html

[7] Specifically in August 2008

[8] http://www.nytimes.com/2008/08/06/business/media/06adco.html?ref=business&$_r = 0$

[9] http://www.google.com/trends/

[10] For example Autos&Vehicles or Arts&Entertainment category

[11] We will use the term GSV (Google search volume) from now on when talking about Google Trends data.

analyzes only a percentage of all searches to determine *"how many searches have been done for the terms you've entered compared to the total number of Google searches done during that time"* [12]. This means that Google Trends randomly chooses a subset of all searches within the same time and location parameters and then computes GSV from this subset.

The process Google Trends performs to compute GSV for a particular keyword was precisely described by Vakrman (2014). Google Trends takes the absolute number of searches for the keyword in time t (day, week, month) and geographic region g, denoted $ASV_{keyword}^{t,g}$ and

1. Divide this $ASV_{keyword}^{t,g}$ by the total number of searches within the same time and geographic region - $ASV_{total}^{t,g}$ - to obtain the relative search volume for the keyword - $RSV_{keyword}^{t,g}$:

$$RSV_{keyword}^{t,g} = \frac{ASV_{keyword}^{t,g}}{ASV_{total}^{t,g}} \tag{1}$$

2. Rescale those relative search volumes over a chosen time interval in such a way that the maximum RSV in this interval gets GSV value equal to 100:

$$GSV_{keyword}^{t,g} = \frac{RSV_{keyword}^{t,g}}{MAX(RSV_{keyword}^{t_0,g}, ..., RSV_{keyword}^{t_T,g})} \cdot 100 \tag{2}$$

   where t_0 and t_T demarcate the specified time interval[13] (Vakrman, 2014, p. 18). Note, however, that these numbers - $ASV_{keyword}^{t,g}$ and $ASV_{total}^{t,g}$ - are obtained from a subset of searches sampled by Google.

Data are available on daily, weekly (Sunday-Saturday) and monthly basis with respect to these specifications: 1) Daily data are accessible only if the selected period is between one to three months[14]. Google updates the daily search volume time series every day, but with the lag of two days. Hence, the latest search volumes available today are the ones from the day before yesterday. 2) Weekly search volumes can be obtained by selecting a period longer than three months. 3) For logged-in users, Google offers the option to

[12] https://support.google.com/trends/?hl=en#

[13] For example, if the entire available period (from 2004 to present) is chosen, data on GSV will appear with a monthly frequency and t_0 and t_T will denote the first and the last month of this period respectively.

[14] It should be noted that Google Trends also allows to acquire daily data for the past seven days.

download data of a daily or weekly frequency into a CSV file. Monthly data are, however, available only in an online chart, and, if the selected period is longer than or equal to three years. Therefore, it is possible to download weekly data for the entire available period at one time, but these data will appear with monthly frequency in the corresponding online chart.

Trends also allows to compare queries: up to five search terms for the same geographic area or up to five geographic areas for a single search term. Therefore, we opted for downloading according to the various groups of five. As noted earlier, it causes a problem, since each series has the maximum GSV equal to 100 and the rest of the series is rescaled. Hence, if someone compares the maximum GSV for a particular keyword in one series in a specified time period with maximum GSV of the same keyword in the identical period but among another series, he or she finds out that these maximum values differ[15]. To overcome this problem, we formed series with one query common to all of them. Search volumes for distinct queries over the same time period were then adjusted to make the maximum GSV for this common query equal in all series. This process can be formally expressed as follows: Let $GSV_1, ..., GSV_N$ be Google search volumes for N (distinct) queries. Let t be a specified time interval (e.g., 1.1.2009-31.3.2009). Then $MAX(GSV_1^t)$ denotes the first query's maximum GSV over the time interval t, ..., $MAX(GSV_N^t)$ the Nth query's maximum over the time interval t. The common query can be any query with at least one non-zero GSV in each series. Let us suppose the query i, where i is between 1 and N, is the common query. Let us denote its maximum GSV in the first series $MAX(GSV_{i,1}^t)$, ..., its maximum GSV in the last series $MAX(GSV_{i,k}^t)$, where k is the number of series. To equalize these maximum values, let us choose e.g. $MAX(GSV_{i,1}^t)$ (i.e., the common query's maximum in the first series) to be the default maximum according to which all other series are adjusted or rescaled. Rescaling is then accomplished by multiplying all GSVs in series $2, ..., k$ by

$$MAX(GSV_{i,1}^t) \cdot \frac{1}{MAX(GSV_{i,j}^t)} \quad \forall j \in 2, ..., k \tag{3}$$

This yields the same maximum values of the common query in all series

[15] The maximum values would be the same if the keyword's $RSV_{keyword}^{t,g}$ over this time period would be the highest among all RSVs in the first group of queries (first series) as well as among all RSVs in the second group. The maximum GSV for the keyword would then equal to 100 in both series.

over the time interval t[16]. As noted by Kristoufek (2013b), this process will probably generate some rounding errors but we assume that these errors are random. Moreover, we can not avoid rounding either, since the data are already rounded by Google.

Google also imposes limitations on its data regarding low search volume and duplicate searches: it excludes rarely searched terms, i.e., terms with search volume under a designated threshold, although the actual value of the threshold is not released by Google. The exclusion rule also applies to repeated searches, i.e., searches submitted to Google by the same person within a short time period[17].

Wikipedia Trends data

Wikipedia Trends[18], as opposed to its Google counterpart, publishes the absolute search volumes for queries realized by English-speaking Wikipedia users. The database works as follows. After entering a keyword of interest, the database shows how many times the corresponding Wikipedia page (if exists) has been requested on each day between 1.1.2008 and the last day of the previous month[19]. More recent data are available via stats.grok.se, a service that also provides data on how often pages on the English-language Wikipedia have been viewed per day, ranging from 10.12.2007 up to the day before yesterday. This database (stats.grok.se), however, allows to download the data only in JSON format and by months. Accordingly, data on Wikipedia pageviews were obtained by accessing www.wikipediatrends.com, the page which enables to download the data (without the obligation to be logged-in) in a CSV file and for the entire available period at once[20].

Data from both services are available also in an online chart, and, are officially referred to as Wikipedia pageviews statistics. Nevertheless, to maintain the convention introduced above, we will use the term Wikipedia search volume or WSV when talking about Wikipedia Trends data. As with

[16] Since we used daily data and these are only accessible in the three-month intervals, we had to rescale all series in the first three-month interval according to the maximum GSV of the common query in one of these series over this interval, then rescale all series in the second three-month interval according to the maximum GSV of the common query in that interval, etc., till we rescaled the data downloaded for the entire analyzed period.

[17] https://support.google.com/trends/?hl=en#

[18] http://www.wikipediatrends.com/

[19] Actually, data from the previous month are available only after the first half of the ongoing month

[20] We downloaded data in the second half of April 2015, thus obtaining pageviews for the period 1.1.2008-31.3.2015

Google, it is possible to enter only 5 search terms into one request field to compare their WSVs, therefore, we downloaded data according to the various groups of five. Wikipedia also offers several data transforming options, e.g., to take the logarithm of WSVs or to adjust them to 0 through 1 range.

3.2 Sample construction

The data set consists of companies that were constituents of the Dow Jones Industrial Average (DJIA) index continuously from 8.4.2004 to 19.3.2015, which totals 22 firms[21].

DJIA is one of the oldest and most watched indexes, which tracks performance of thirty American blue chips, i.e., stocks of the top-performing publicly traded companies[22]. This assures that data on search frequency for company-related queries will not exhibit high percentage of zero Google search volumes due to not reaching the limit value[23].

Following Kristoufek (2013b), with the difference that data of a daily frequency are employed[24], we focus on GSV of stock ticker symbol (e.g., AXP for American Express Company and INTC for Intel Corporation) as well as on GSV for the combination of the word "stock" and the ticker symbol ("stock AXP", "stock INTC"). As far as I am concerned, the authors of Da et al. (2011) were the first to use ticker symbols to capture the interest in particular stocks. Another option, which was employed by Bank et al. (2011) and Vlastakis and Markellos (2012) among others, is to search for a unique search term associated with the company name (e.g., "Disney" for Walt Disney Company). The authors claim that this way it is possible to measure information demand from a much broader audience, not only from professional market participants (Bank et al., 2011). Da et al. (2011), however, argues that "investors may search the company name for reasons unrelated to investing" (Da et al., 2011, p. 1466), thus including irrelevant components to the search volume time series. Logically, if one searches for "Boeing", he or she can be interested in its new airplane model rather than in financial information about that company. The problem is even more apparent if the company name has multiple meanings, e.g., "Apple" or "Siemens". Moreover,

[21] We will refer to these twenty-two companies from now on as to the DJIA companies or DJIA constituents, even if it is not exactly correct.

[22] The DJIA data set was also used by Vlastakis and Markellos (2012), Kristoufek (2013b) and Preis (2013) among others.

[23] See Section 3.1 for the discussion of Google's treshold

[24] Kristoufek (2013b) used weekly data

the exact term investors use when searching for financial information about the company can substantially vary, but hardly anyone uses a full company name (e.g., "Wal-Mart Stores, Inc."). On the other hand, a ticker symbol is always unequivocally associated with the firm. Thus a person searching for MSFT on Google is probably interested in financial information about the stock of Microsoft (Da et al., 2011). For these reasons, if we want to assess the effect of investors' interest in a company's stock, it is arguably not inconvenient to use ticker symbol for the identification of the stock in Google.

We are aware of any ticker with generic meaning, such as CAT, DIS, HD. *GSV* for these tickers is usually very high, but it hardly captures the investors' interest in the corresponding company's stock. Therefore, we manually go through the list of tickers and exclude the ones which are easily interchangeable with other stock unrelated terms and abbreviations. The issue with noisy tickers is obviously not present if we consider searching for the term "stock" along with its ticker symbol, e. g., "stock CAT".

Next problem concerns Google's treatment of unpopular keywords, i.e., keywords with insufficient search volume. As previously noted, Google imposes a limitation on its data regarding search volume, and therefore queries with search volume under a designated threshold are automatically assigned zero *GSV*. These omitted *GSV*s may not be truly zero[25], therefore it causes a problem when someone wants to plug in zero value instead. Another option is to assign to these *GSV*s a random number below the truncation threshold, although, as Vakrman (2014) in his master thesis argues, *"it would introduce a synthetic volatility and trend to the GSV time series that may be different from the actual volatility and trend in it"* (Vakrman, 2014, p. 21). Arguably the most appropriate solution, when examining the effect of changing *GSV*s, is to consider only the valid *GSV*s, which reduces sample size and also selectively drops the lowest *GSV*s but does not bring any synthetic information to our data[26] (Vakrman, 2014).

For these reasons, we excluded tickers with infrequent queries, and, as noted earlier, tickers with ambiguous meanings, namely AXP, BA, CAT, DD, DIS, HD, KO, MCD, PG, UTX and VZ. For the second - stock/ticker - approach, we had to remove observations for AXP, DD, MCD, MMM, MRK,

[25] Actually, it is very improbable that *GSV* for any of these stock-related terms is truly zero.

[26] This truncation issue is rather unpleasant for computing portfolio weights; we will address this issue in Section 4.

PFE, UTX due to infrequent search queries.

We utilized search volumes from Google for these two approaches in two different periods (due to data availability; GSVs for the dates before these periods are frequently missing due to not reaching the limit value) - 1.1.2007-13.3.2015 for the ticker approach and 1.1.2013-19.3.2015 for the stock/ticker one.

To quantify the interest in a stock, we also utilized pageviews of the online encyclopedia Wikipedia, which are available in its English version. These pageviews register how many times a specific article has been viewed on a particular day. Wikipedia's advantage over Google is that Wikipedia can distinguish between searches for e.g. apple the fruit and Apple the company, because every Wikipedia page has an unique title unequivocally associated with the page's content. Therefore, we are able to separate views of the page named "Apple" (i.e. the page about apple the fruit) from the views of "Apple Inc.". This enables us to include all DJIA companies[27] to the analysis by downloading the data on Wikipedia pageviews of articles dedicated exactly to these companies, the searched keyword being the article's title (e.g. "Microsoft" for Microsoft Corporation, "3M" for the 3M Company). Table 1 lists the companies in DJIA sample along with the corresponding stock tickers and Wikipedia page titles.

Contrary to Google, Wikipedia imposes no minimum-searches threshold and publishes the actual number of its pageviews; hence the data were available for the entire period (1.1.2008-31.3.2015)[28]. We did not face the problem with infrequent queries either, again due to the fact that Wikipedia publishes the absolute number of pageviews, and we had chosen to analyze the well-known companies (DJIA constituents). Accordingly, all observations in the analyzed period were non-zero.

Lastly, data on historical prices for DJIA stocks included in the analysis as well as on historical prices for DJIA index come from Yahoo! Finance database[29]. All prices are in US dollars.

[27] Companies that were constituents of the DJIA index continuously between 8.4.2004 and 19.3.2015 and create the initial dataset.

[28] As noted earlier, Wikipedia Trends database ranges back to the beginning of 2008.

[29] http://finance.yahoo.com/

4 Methodology

This section describes the construction of portfolio diversification strategy and the process of evaluating the performance of portfolios created in line with the proposed strategy.

Since this thesis strives to deal with the utility of web search data in portfolio selection and risk diversification, we consider deriving portfolio weights based on the assumption that the surge in company-related web search volume negatively affects price volatility of the company's stock, and then examine performance of such portfolios.

The positive interdependence between web search volume and stock price volatility was empirically confirmed by Dimpfl and Jank (2011), Chen (2011) and Vlastakis and Markellos (2012) among others[30]. We assume that an increase in web searches for a keyword related to the firm can subsequently cause the corresponding stock price to be more volatile, thus increases the risk of investing in the stock. Conversely, less popular stocks should be also less risky. Accordingly, we are interested in a diversification strategy that discriminates popular stocks (and prefer less popular ones), where the popularity is measured by the extent to which people search on Google or Wikipedia for stock-related terms. Such diversification strategy is then utilized in the construction of portfolios to decrease their total riskiness.

Since we focus on search volume data with daily frequency, we construct a series of daily returns for each stock employed in the analysis to match this pattern:

$$r_{i,t} = \frac{p_{i,t}^{adjClose} - p_{i,t-1}^{adjClose}}{p_{i,t-1}^{adjClose}}, \tag{4}$$

where $r_{i,t}$ stands for the daily return of stock i on day t and $p_{i,t}^{adjClose}, p_{i,t-1}^{adjClose}$ are close prices adjusted for dividends and splits provided by Yahoo! Finance.

Diversification strategy is constructed according to Kristoufek (2013b), with the only difference that data of a daily frequency are employed and Wikipedia searches are utilized along with those from Google[31]. Kristoufek (2013b) proposed a novel approach to portfolio diversification based on the search volumes of stock-related terms. The diversification strategy stems in an idea that the more popular stocks should be discriminated by assigning them lower weights in the final portfolio. Less popular stocks, on the other

[30] See Section 2 for more details.

[31] Kristoufek (2013b) used Google search volume of a weekly frequency.

hand, should be preferred. Let $V_{i,t}$ be a search volume (Wikipedia or Google) for a keyword related to stock i on day t. Portfolio weight of stock i on day t is defined as:

$$w_{i,t} = \frac{V_{i,t}^{-\alpha}}{\sum_{j=1}^{N} V_{j,t}^{-\alpha}}, \tag{5}$$

where α is *"...a power-law parameter measuring the strength of discrimination of a stock being too frequently searched for"* (Kristoufek, 2013b, p. 2) and N is the number of stocks in the portfolio. Adding $\sum_{j=1}^{N} V_{j,t}^{-\alpha}$ into the equation ensures that portfolio weights on each particular day sum up to one[32]: $\sum_{i=1}^{N} w_{i,t} = 1$ for all t.

The value of α indicates to which extent popular stocks are discriminated in the portfolio as well as whether they are actually discriminated: if α is positive, the more popular stocks are assigned lower weights in the portfolio, if α is negative, the opposite is true. For $\alpha = 0$, a uniformly distributed portfolio with $w_{i,t} = w = \frac{1}{N}$ for all i and t is obtained.

The proposed diversification strategy ensures that each stock with a non-zero search volume is at least marginally present in the portfolio. Hence, even if the popular stocks are discriminated (i.e., the case when α is positive), their weights in the portfolio do not vanish too quickly and frequently with increasing positive α, which would happen if we employed, e.g., an exponential discrimination rule (Kristoufek, 2013b, p. 2).

It should be noted that Google search volumes for some stock-related queries were not available due to not reaching the limit value[33]. This is arguably the biggest shortcoming of Google Trends data because these queries have most probably some positive (but low) search volumes that we cannot observe. Accordingly[34], it is incorrect to treat these GSVs as if they were zero and plugging a zero-value search volume into the Equation for computing portfolio weights (Equation 5) would not be feasible either, at least for non-negative α. Therefore, we deal with these missing GSVs in the following way: if the (Google) search volume $V_{i,t}$ of stock i on day t is not available, we exclude this stock from the portfolio on day t by assigning it a zero weight in the portfolio - $w_{i,t} = 0$. However, assigning to the stock with a zero stock-related search volume a zero portfolio weight, goes against the

[32] Provided that at least one (out of N) search volume $V_{i,t}$ is available or non-zero for each t.

[33] As mentioned in Section 3.2 Google assigns a zero search volume to queries with a search volume under a given threshold. However, the actual value of the threshold is not released by Google.

[34] Consult Section 3.2.

main principle of the proposed diversification strategy: to prefer less popular stocks in the portfolio. Fortunately, the issue of missing search volumes is not present when Wikipedia searches are considered, since Wikipedia publishes the absolute number of pageviews. Hence, data on pageviews[35] were available (non-zero) for all the DJIA companies included in the analysis[36].

We focus on the in-sample and out-of-sample performance of portfolios constructed in line with proposed methodology. As Kristoufek (2013b) argues, *"The former is a standard approach to measure quality of portfolio optimization but the latter is more useful for practical applicability of the diversification strategy."* (Kristoufek, 2013b, p. 2). The in-sample performance is based on portfolio weights rebalancing each trading day according to Equation 5 and realizing gains/losses on the same day. Therefore, the in-sample portfolio return on day t is defined as:

$$R_t^{IS} = \sum_{i=1}^{N} w_{i,t} \cdot r_{i,t} \tag{6}$$

where $w_{i,t}$ is portfolio weight of stock i on day t derived according to Equation 5 and $r_{i,t}$ is daily return of the corresponding stock based on Equation 4. Again, N is the number of stocks in the portfolio.

The out-of-sample stems in rebalancing portfolio on day t (if this is the trading day) using the most recent Google search volumes, i.e., the volumes from day $t-2$. This is because Google Trends and Wikipedia[37] publishes daily search volumes with the lag of two days, hence the latest search volumes available today are the ones from the day before yesterday. The out-of-sample portfolio return on day t can therefore be expressed as:

$$R_t^{OS} = \sum_{i=1}^{N} w_{i,t-2} \cdot r_{i,t} \tag{7}$$

where $w_{i,t-2}$ is portfolio weight of stock i on day $t-2$ derived according to Equation 5 and $r_{i,t}$ is daily return of the same stock on day t defined in Equation 4.

Portfolio performance for both in- and out-of-sample is examined by means of standard deviation and the Sharpe ratio. Standard deviation measures the variability of returns of a portfolio; the more the portfolio returns

[35] As noted earlier, we downloaded the data on pageviews of articles dedicated exactly to the DJIA companies.

[36] See Table 1

[37] Through http://stats.grok.se/.

are dispersed the higher the risk of the portfolio. Sharpe ratio is the most widely used method for evaluating risk-adjusted return[38]. It is also referred to as *"the standardized average return of the portfolio"* (Kristoufek, 2013b, p. 2) and is defined as the ratio between average return and standard deviation. In line with financial literature, we are interested in the Sharpe ratio maximizing portfolios as well as in portfolios with minimum risk (i.e. minimum standard deviation).

Stocks included in the portfolio as well as the corresponding portfolio weights are retrieved with respect to four different scenarios depending on whether Google or Wikipedia searches are utilized in calculating portfolio weights[39] and which search query is used for the identification of a stock on the internet.

Searching for the ticker symbol on Google

First, we focus on search frequency on Google with the search term being the stock's formal ticker symbol (e.g. GE for General Electric Company). In line with Section 3.2 we had to omit tickers with ambiguous meanings and infrequent queries[40], ending-up with the sample of 11 stocks. Although the data on Google search volumes are accessible from the beginning of 2004, we analyzed a period of 1.1.2007-13.3.2015 (2994 days, from which 931 are non-trading days) due to the significant number of missing queries before the year 2007.

Google search volume for the combination of the word "stock" and the ticker symbol

Next, we take Google search volume and, as for the search term, the combination of the word "stock" and the ticker symbol (e.g. "stock IBM") to ensure that the corresponding search frequency is unequivocally associated with interest in the company's stock. However, data were available only from 1.1.2013 to 19.3.2015 (808 days out of which 251 are non-trading days) due to infrequent search queries before the analyzed period. The final portfolio counts 15 stocks; for the rest, Google search volumes were still frequently missing[41].

[38] http://www.investopedia.com/terms/s/sharperatio.asp

[39] See Eq. 5

[40] Tickers AXP, BA, CAT, DD, DIS, HD, KO, MCD, PG, UTX and VZ were removed.

[41] Stocks AXP, DD, MCD, MMM, MRK, PFE, UTX were excluded from the analysis.

Wikipedia pageviews

Now, to capture interest in particular stocks, we focus on Wikipedia pageviews of articles dedicated exactly to the DJIA companies, the searched keyword being the (unique) article's title. This way it is possible to include all DJIA constituents[42] to the analysis, because we can exactly separate the views of articles about these companies from the views of the other articles. Since Wikipedia 1) publishes the absolute search volume for queries submitted by Wikipedia users and 2) does not impose any truncation threshold[43], we do not have to deal with missing (or zero) search volumes. Thus the entire available period - 1.1.2008-31.3.2015 (2647 days from which 823 are non-trading ones) - can be analyzed.

Combination of Google and Wikipedia searches

Lastly, queries from Google and Wikipedia are utilized together for calculating portfolio weights. The weight $w_{i,t}^{comb}$ (where *comb* stands for the word "combined") of stock i on day t is defined as the arithmetical average of the weight of stock i on day t derived using Google search volume and the same stock's weight on the same day derived using Wikipedia search volume (Wikipedia pageviews):

$$w_{i,t}^{comb} = \frac{w_{i,t}^{G} + w_{i,t}^{W}}{2} \tag{8}$$

where $w_{i,t}^{G}$ and $w_{i,t}^{W}$ are portfolio weights calculated from Eq. 5 using Google and Wikipedia search volumes respectively.

We focus on Google searches for the ticker symbol (with the total of 11 stocks analyzed) as well as on Wikipedia pageviews of articles dedicated to these 11 companies. Our portfolio therefore contains stocks that create the portfolio when utilizing Google searches for tickers (the first approach): GE, IBM, INTC, JNJ, JPM, MMM, MRK, MSFT, PFE, WMT, XOM. Wikipedia weights entering Equation 8 had to be recalculated according to Equation 5 using search volumes only for these eleven stocks to ensure that $\sum_{i=1}^{N} w_{i,t}^{comb} = 1$ for all t. The analyzed period is from 1.1.2008 to 13.3.2015[44] (2629 days from which 817 are non-trading days).

[42] See Table 1

[43] See Section 3.2 for details.

[44] This period is the intersection of the two periods - the period over which the utility of Google searches for ticker symbols in portfolio optimization is examined and the period used for analyzing Wikipedia searches. See above.

5 Results

This section presents the results for all scenarios specified in Section 4. It is organized as follows. For each scenario, we first depict standard deviations and Sharpe ratios obtained for the search-based portfolios, both in-sample and out-of-sample. For reference, we also show the performance of the DJIA index. Second, we discuss the implications of these results and provide a comparison of the evolution of the Sharpe ratio maximizing portfolios with the evolution of the DJIA index.

5.1 Searching for the ticker symbol on Google

This part evaluates the performance of portfolios constructed using Google search volumes, with the search query being the stock's formal ticker symbol.

In Figure 1 we present standard deviations for portfolios, both for the out-of-sample and in-sample strategies[45] For illustration, we also plot standard deviation of the DJIA index. The discrimination parameter, α, always varies between -2 and 2 with a step of 0.1. The behavior of the standard deviation is essentially the same for both the in-sample and out-of-sample: it decreases with increasing α, reaching its minimum at $\alpha = 2$. Although the in-sample standard deviation is lower than the out-of-sample standard deviation for almost all the values of the discrimination parameter (except for α between -0.7 and -0.1). The search-based Sharpe ratios are shown in Figure 2. Considering the in-sample approach, the ratio decreases with increasing α, up to $\alpha = -1.2$, and then steadily grows, reaching its maximum at $\alpha = 2$. For the out-of-sample, the ratio first slightly declines (between $\alpha = -2$ and $\alpha = -1.3$), and then rises up to its maximum at $\alpha = 0.4$. As in the case of standard deviation, the in-sample approach outperforms the out-of-sample one, reaching the higher Sharpe ratio.

These results suggest that the strategies based on Google search volume for ticker symbols are able to reach lower risk level than the uniformly weighted portfolio (i.e. the case when $\alpha = 0$) and, interestingly, even than the passive buy-and-hold strategy (buying the DJIA at the beginning of 2007, holding and selling it in the first half of March 2015). Search-based strategies also outperform both the benchmark DJIA index and the uniformly weighted portfolio with respect to the Sharpe ratio - approximately 0.032 Sharpe ratio of the out-of-sample performance compared to less than

[45] The strategies are defined in Section 4.

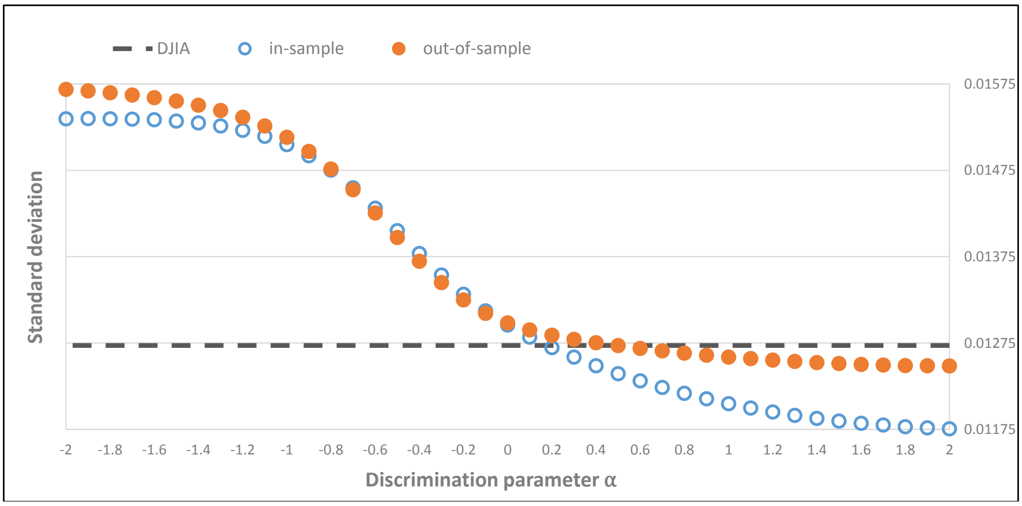

Figure 1: **Standard deviations for portfolios based on ticker symbols.** The figure shows standard deviations for portfolios constructed using Google search volume for ticker symbols. Both in-sample (empty blue circles) and out-of-sample (full orange circles) performances are shown in the figure. The discrimination parameter α ranges between -2 and 2 with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA index. Standard deviation for both the in-sample and out-of-sample strategies decreases with increasing α, reaching the minimum at $\alpha = 2$.

0.02 of the DJIA index. Moreover, regardless of the value of α (i.e. discriminating or preferring more popular stocks in the portfolio), both strategies reach higher Sharpe ratio than the DJIA strategy, which is something rather unexpected. According to our predictions, we expect higher Sharpe ratio only for (a few) portfolios based on positive α, meaning that the strategy discriminating more searched stocks leads to the higher Sharpe ratio than the DJIA strategy.

It should be emphasized that the query-based strategy is successful even in the out-of-sample, reaching lower risk level and higher Sharpe ratio than the uniformly weighted portfolio and even than the passive buy-and-hold strategy. The out-of-sample performance is inferior to the in-sample performance, which is not unexpected. And finally, the diversification strategies based on positive α strongly outperform the ones with negative α, implying that discrimination of popular stocks pays off in the portfolio selection.

For illustration, Figure 3 provides us with the evolution of portfolio value. The in-sample and out-of-sample evolutions are shown for the discrimination parameter α which maximizes the Sharpe ratio[46]. For comparison, we

[46] Considering this - ticker - strategy, $\alpha = 2$ for the in-sample and $\alpha = 0.4$ for the out-of-sample.

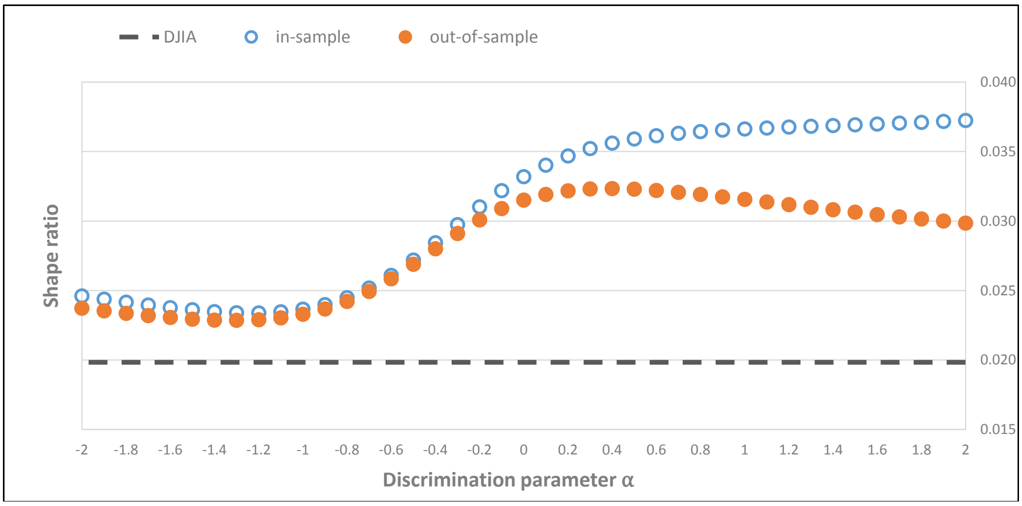

Figure 2: **Sharpe ratios for portfolios based on ticker symbols.** The figure displays the behavior of the Sharpe ratio for portfolios constructed using Google search volume for ticker symbols. Both in-sample (empty blue circles) and out-of-sample (full orange circles) performances are shown in the figure. The discrimination parameter α ranges between -2 and 2 with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA index. The behavior of the Sharpe ratio differs for the two approaches: for the in-sample, the maximum ratio is obtained for $\alpha = 2$, for the out-of-sample, the ratio is maximized for $\alpha = 0.4$.

also plot the evolution of the DJIA index in the analyzed period (1.1.2007-13.3.2015). It should be noted that these portfolios are not the profit maximizing ones but only the Sharpe ratio maximizing ones, since this thesis focus on the utility of web search data in portfolio diversification and does not strive to find the profit maximizing strategy.

For the in-sample, the portfolio value at the end of the analyzed period is approximately 190% of its initial value, therefore reaching a cumulative profit of 90%. The out-of-sample approach is slightly worse, with a cumulative profit of 85%. We are mainly interested in the comparison of the out-of-sample strategy and the DJIA strategy as it measures how much would the search-based be profitable compared to the passive buy-and-hold strategy. According to the Figure 3, the passive buy-and-hold strategy reaches a cumulative profit of 52%. This indicates that the search-based strategies dominate the DJIA strategy but not nearly as strongly as in the case of Kristoufek (2013b), who obtained more than a quadruple profit of the ticker strategy with respect to the DJIA strategy. Note, however, that Kristoufek (2013b) used weekly data and examined a different period. Moreover, the results show a huge loss after the beginning of 2009, obviously due to the financial crisis which was greatly intensified by the Lehman Brothers bankruptcy in September 2008. The value of the search-based portfolios had not slumped

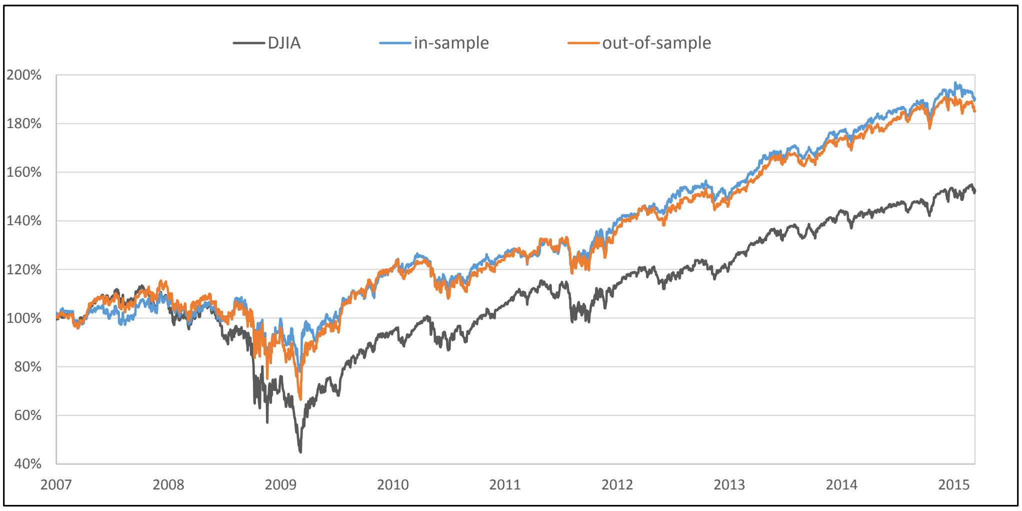

Figure 3: **Evolution of portfolios based on ticker approach.** Black line represents the evolution of the DJIA index, blue and orange lines show the development of the in-sample and out-of-sample strategies, respectively. Evolution is shown for the values of α which maximizes the Sharpe ratio for both the search-based strategies. For practical purposes, the comparison between the out-of-sample and the DJIA strategy is essential. The out-of-sample reached the cumulative profit of 185% compared to the DJIA cumulative profit of 152%.

so much as the value of the DJIA index, although these portfolios had not been able to forestall the loss anyway, which is in line with the results of Kristoufek (2013b).

To conclude, the ticker strategies were able to reach higher cumulative profit than the passive buy-and-hold strategy. However, as Kristoufek (2013b) pointed out, the active (search-based) strategies need to yield an extra profit to overcome the (transaction) costs, which raises from daily (or weekly) realigning the portfolio weights. In this case, the strategy based on rebalancing the portfolio each day can bring substantial costs, thus is required to reach a notably larger cumulative profit than the passive DJIA strategy. And the difference in cumulative profits (between the out-of-sample and the DJIA strategy) of 33% over more than 8 years is not high enough to overcome these costs. Kristoufek (2013b), using the data of a weekly frequency, reached a difference of 125% over 8.5 years, which is much more satisfying.

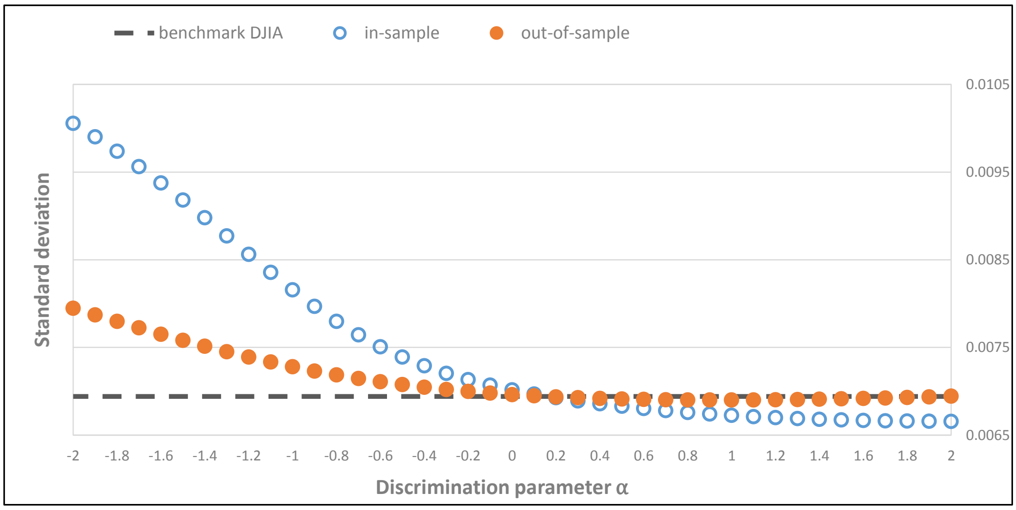

Figure 4: **Standard deviations for the stock/ticker strategy.** The figure shows standard deviations for portfolios constructed using Google search volume for ticker symbols combined with the word “stock”. Both in-sample (empty blue circles) and out-of-sample (full orange circles) performances are shown in the figure. The discrimination parameter α ranges between -2 and 2 with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA strategy. Minimum variance portfolios are located at $\alpha = 2$ (in-sample) and $\alpha = 1$ (out-of-sample).

5.2 Google search volume for the combination of the word “stock” and the ticker symbol

In this subsection, we present and discuss the results for portfolios based on Google search queries for the combination of the word “stock” and the ticker symbol.

Figure 4 shows standard deviation for the in-sample and out-of-sample portfolios based on the stock/ticker approach. Standard deviation for the in-sample first sharply decreases up to $\alpha = -0.4$ where the declining gets slower, but still continues, till the deviation reaches its minimum at $\alpha = 2$. For the out-of-sample, the standard deviation decreases with increasing α between $\alpha = -2$ and $\alpha = 1$ where it reaches the minimum, and for $\alpha > 1$, it slowly growths. Note that the range in which the values of the in-sample standard deviation can be found is noticeably wider than the one in which the values of the out-of-sample standard deviation are located.

Sharpe ratios obtained for the search-based strategies (both in-sample and out-of-sample) are displayed in Figure 5. The behavior of the ratio in-sample is comparable to its out-of-sample development: it climbs steadily up to $\alpha = 2$ where the maximum Sharpe ratios for both the in-sample and out-of-sample performances are located.

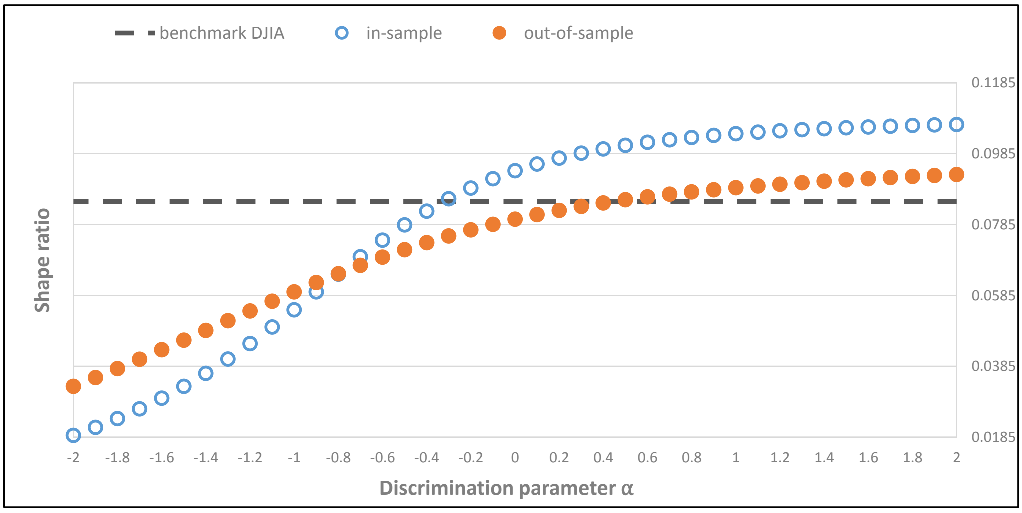

Figure 5: **Sharpe ratios for the stock/ticker strategy.** The figure displays the behavior of the Sharpe ratio for portfolios constructed using Google search volume for ticker symbols combined with the word "stock". Both in-sample (empty blue circles) and out-of-sample (full orange circles)performances are shown in the figure. The discrimination parameter α ranges between -2 and 2with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA strategy. The in-sample and out-of-sample performance of the Sharpe ratio is comparable: it increases with increasing α, reaching the maximum at $\alpha = 2$.

As previously, the Google Trends based strategies dominate both the uniformly weighted portfolio and the benchmark DJIA index. However, the dominance is not as apparent as in the previous case - the out-of-sample Sharpe ratio is approximately 9% higher than the DJIA Sharpe ratio compared to the 63% difference in the previous case. Moreover, using the search-based strategies, we again obtain portfolios with lower risk than both the uniformly weighted portfolio and the benchmark DJIA index. As with the strategies based solely on ticker symbols, the stock/ticker approach is also successful in the out-of-sample, yielding portfolios with higher standardized returns and lower standard deviations than the passive buy-and-hold strategy. Albeit the in-sample performance is again superior to the out-of-sample performance. These results also confirm that portfolios with positive α strongly outperform those with negative α, suggesting that the diversification strategy based on the discrimination of popular stocks pays off in the final portfolio.

The evolution of the Sharpe ratio maximizing portfolios along with the development of the DJIA strategy is provided in Figure 6. Here, the values of the search-based portfolios practically track the DJIA index, reaching the cumulative profit of approximately 36% (for both the in-sample and

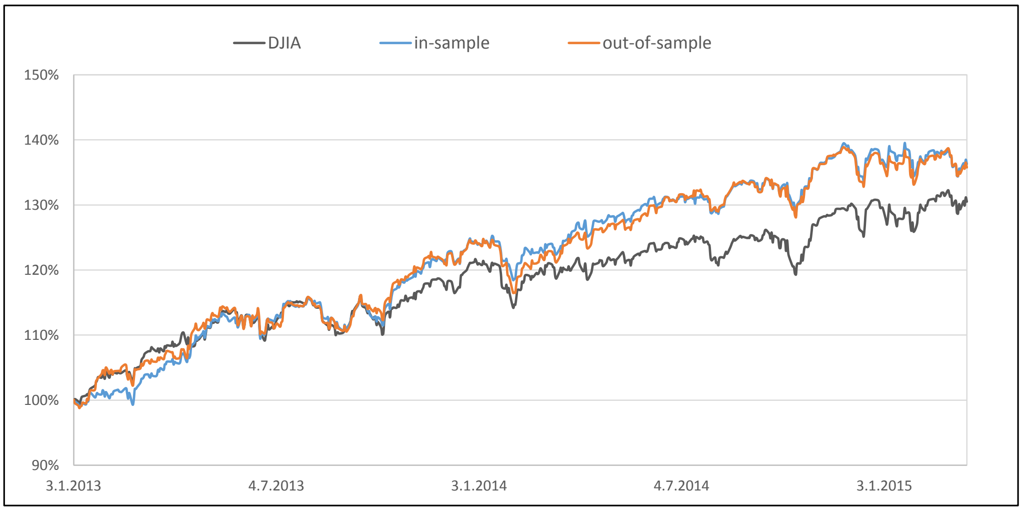

Figure 6: **Evolution of portfolios based on stock/ticker approach.** Black line represents the evolution of the DJIA index, blue and orange lines show the development of the in-sample and out-of-sample strategies, respectively. Evolution is shown for the values of α which maximizes the Sharpe ratio for both the search-based strategies. For practical purposes, the comparison between the out-of-sample and the DJIA strategy is essential. The cumulative profit of the search-based strategies is 36% higher than the one of the passive DJIA strategy.

out-of-sample strategies). The DJIA strategy, with a cumulative profit of approximately 30%, is just slightly worse. In line with the discussion about the previous results, the difference of 6% over more than 2 years would not be able to overcome the costs related to the active search-based strategy, which stems in daily rebalancing the portfolio. The small difference in cumulative profits is even more observable than for the ticker strategy, because in this case the values of the DJIA index are very close to the ones of the search-based portfolios. This essentially confirms the stock/ticker results of Kristoufek(2013b).

5.3 Wikipedia pageviews

Here we show the in-sample and out-of-sample performances of Wiki-based strategies, with the search query being the title of the company's page on Wikipedia.

Figure 7 shows standard deviations for portfolios rebalanced according to Eq. 5 using Wikipedia search volumes suggesting how many times the company's page has been requested. Both in-sample and out-of-sample performances are depicted in the figure. The behavior of the standard deviation is rather unexpected: it firstly decreases with increasing α, which is in line

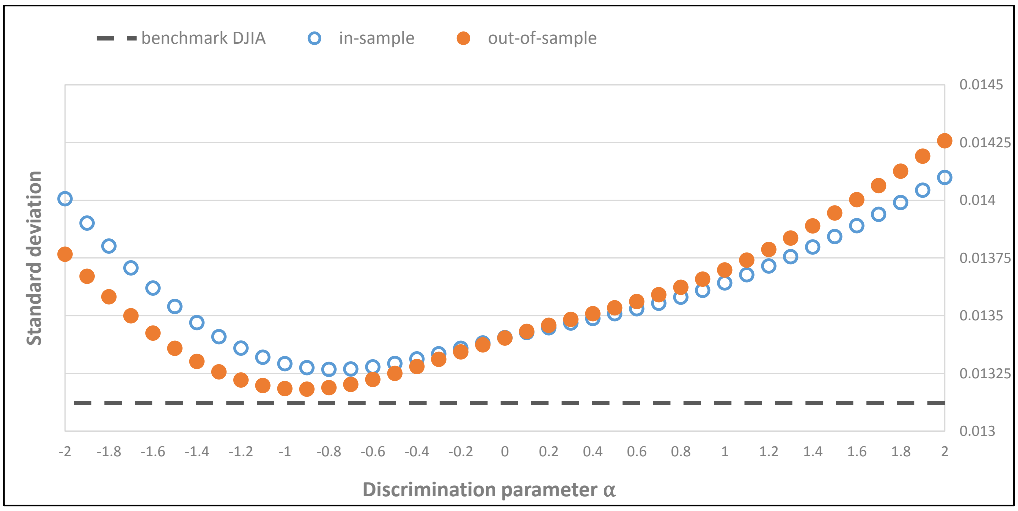

Figure 7: **Standard deviations for Wikipedia strategy.** The figure shows standard deviations for portfolios constructed by utilizing Wikipedia pageviews. Both in-sample (empty blue circles) and out-of-sample (full orange circles) performances are shown in the figure. The discrimination parameter α ranges between -2 and 2 with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA strategy. Standard deviation for both approaches decreases with increasing α, reaching the minimum at $\alpha = -0.8$ (in-sample) and $\alpha = -0.9$ (out-of-sample), and then increases up to $\alpha = 2$.

with our predictions, although it reaches the minimum at $\alpha = -0.8$ (in-sample) and $\alpha = -0.9$ (out-of-sample) and then increases up to $\alpha = 2$. This implies that the minimum variance portfolios were obtained for such values of the discrimination parameter which prefer more popular stocks in the portfolio.

On the other hand, Sharpe ratio's behavior (depicted in Figure 8) reveals the superiority of the search-based strategies over the passive DJIA strategy: Sharpe ratios for all the values of α are well-above the benchmark DJIA Sharpe ratio. Moreover, they continuously growth for both approaches, reaching the maximum at $\alpha = 1.6$ (in-sample) and $\alpha = 1.5$ (out-of-sample).

The results for Wiki-based strategies slightly differs from what we have previously observed. Albeit the search-based strategies again outperform both the uniformly weighted portfolio and the benchmark DJIA index (and the out-of-sample approach reaches more than a twofold Sharpe ratio than the DJIA index), the lowest risk level is obtained for portfolios constructed using negative α. That implies that portfolios preferring Wiki-popular stocks are more successful in minimizing the standard deviation than those discriminating Wiki-popular stocks. Such behavior of standard deviation contradicts

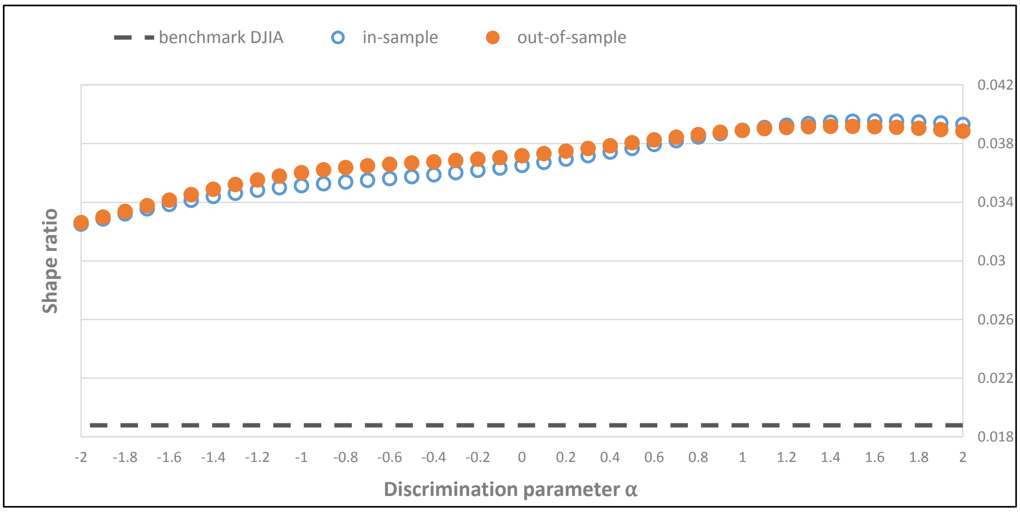

Figure 8: **Sharpe ratios for Wikipedia strategy.** The figure displays the behavior of the Sharpe ratio for portfolios constructed by utilizing Wikipedia pageviews. Both in-sample (empty blue circles) and out-of-sample (full orange circles) performances are shown in the figure. The discrimination parameter α ranges between -2 and 2with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA strategy. The maximum ratios are located at $\alpha = 1.6$ (in-sample) and $\alpha = 1.5$ (out-of-sample).

our model predictions.

Regarding Wiki-based strategies, the out-of-sample performance is not inferior to the in-sample performance (the pattern we have seen previously). This would support our conjectures about the utility of the search-based strategies in portfolio optimization, yet, surprisingly, the out-of-sample strategy outperforms the in-sample one for the discrimination parameters which prefer Wiki-popular stocks in the portfolio. These results suggest that searching on Wikipedia for company-related information may not be a commonly used practice for more sophisticated investors. Wikipedia serves for many people as an important source of general information about various subjects of interest but it is hardly a primary service they would search on for financial information. Finally, we can conclude that Wikipedia search-based strategies significantly outperform the DJIA strategy considering the Sharpe ratio, but they are not able to reach lower risk level.

Figure 9 shows the evolution of Wiki-based portfolios (only the Sharpe ratio maximizing ones) with respect to the evolution of the DJIA index. We can see that the portfolio value for both the in-sample and out-of-sample diversification approaches has doubled during the analyzed period, while the value of the DJIA index increased by about 46%. Therefore, the out-of-sample portfolio (whose comparison with the DJIA index is most essential)

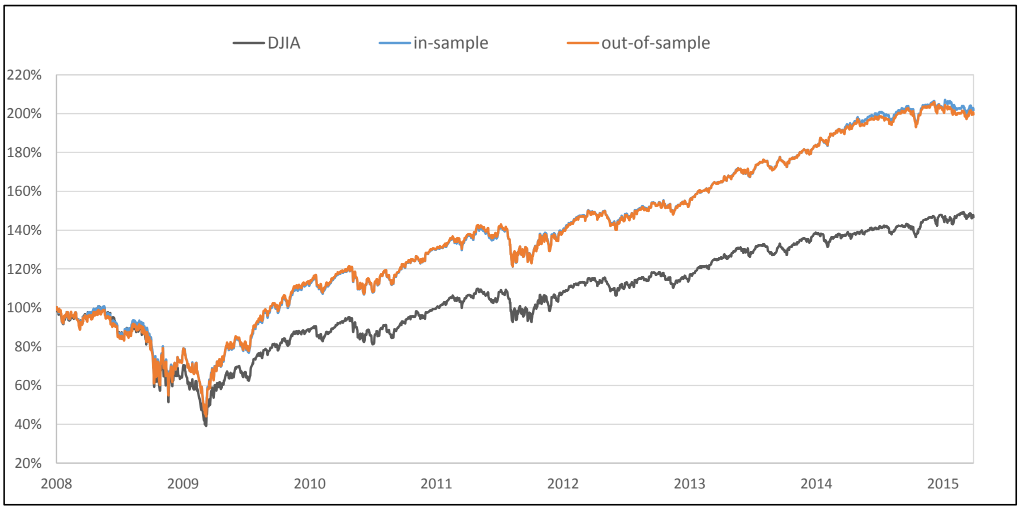

Figure 9: **Evolution of portfolios based on Wikipedia strategies.** Black line represents the evolution of the DJIA index, blue and orange lines show the development of the in-sample and out-of-sample strategies, respectively. Evolution is shown for the values of α which maximizes the Sharpe ratio for both the search-based strategies. For practical purposes, the comparison between the out-of-sample and the DJIA strategy is essential. The out-of-sample outperforms the DJIA strategy by reaching the higher cumulative profit of approximately 54%.

yields the cumulative profit of approximately 100% over 7 years compared to the DJIA cumulative profit of 46%. This is the higher difference than that obtained for both the ticker and the stock/ticker strategies.

5.4 Combination of Google and Wikipedia searches

Lastly, we are keen on evaluating the performance of the diversification strategy in which the search data both from Google and Wikipedia were employed to measure the interest in a company.

We are interested in standard deviations and the Sharpe ratios of portfolios constructed according to this - “combined” - strategy. The strategy stems in utilizing Google search queries along with those from Wikipedia for one selection criterion[47]. Figure 10 provides us with the behavior of standard deviation for this “Combined” strategy. For the in-sample, the deviation declines with increasing α up to $\alpha = 1.9$, where the deviation reaches the minimum. Considering the out-of-sample performance, the deviation decreases between $\alpha = -2$ and $\alpha = 1.6$ where the minimum is located, and for $\alpha > 1.6$, it slightly increases.

Sharpe ratios are shown in Figure 11. Sharpe ratio for the in-sample first

[47] See Section 4 for details

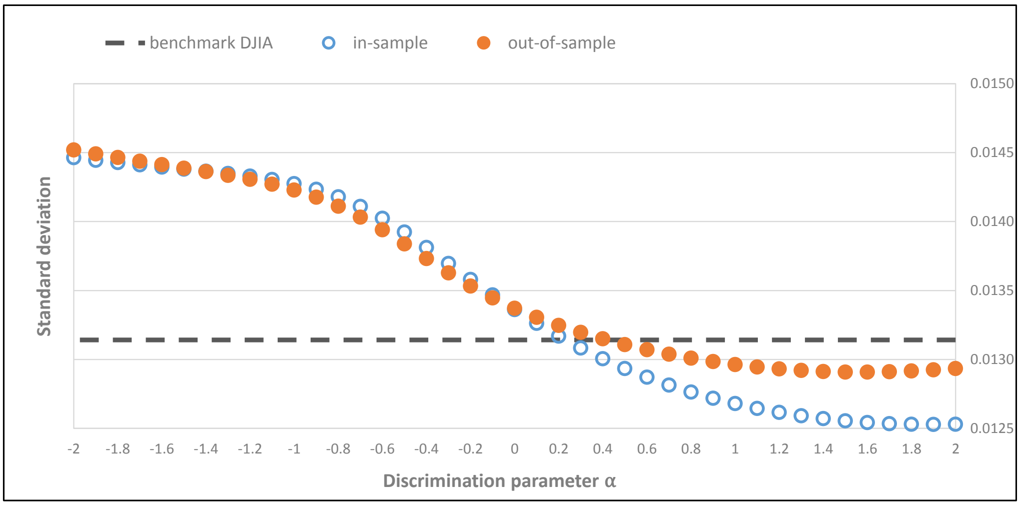

Figure 10: **Standard deviations for combined strategy.** The figure shows standard deviations for portfolios constructed by utilizing Google and Wikipedia searches together in portfolio optimization. Both in-sample (empty blue circles) and out-of-sample (full orange circles) performances are shown in the figure. The discrimination parameter α ranges between -2 and 2 with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA strategy. Portfolios with minimum standard deviation are obtained for $\alpha = 1.9$ (in-sample) and $\alpha = 1.6$ (out-of-sample).

fractionally declines between $\alpha = -2$ and $\alpha = -1.2$, and for $\alpha > -1.2$, it increases up to $\alpha = 2$. For the out-of-sample, the Sharpe ratio also negligibly declines between $\alpha = -2$ and $\alpha = -1.3$, but for $\alpha > -1.3$, it rises up to $\alpha = 1$ where it reaches the maximum.

Strategies based on the combined measure of attention reach lower risk level than the uniformly weighted portfolio and than the passive DJIA strategy. These strategies also dominate the benchmark DJIA and the uniformly weighted portfolio regarding the Sharpe ratio, with the in-sample outperforming the DJIA strategy strongly: 0.038 Sharpe ratio compared to 0.019 of the DJIA index. Even the out-of-sample yields notably higher Sharpe ratio (approximately 0.033) than the passive buy-and-hold strategy, implying that the strategies are successful even in the out-of-sample. Moreover, strategies with positive α again dominate those with negative α, thus confirming our conjectures.

For illustration, we again provide (in Figure 12) the in-sample and the out-of-sample evolution of the Sharpe ratio maximizing portfolios in comparison with the evolution of the DJIA index. The value of the in-sample portfolio at the end of the analyzed period is 187%, which corresponds to the cumulative profit of 87%. This is better than the out-of-sample cumulative

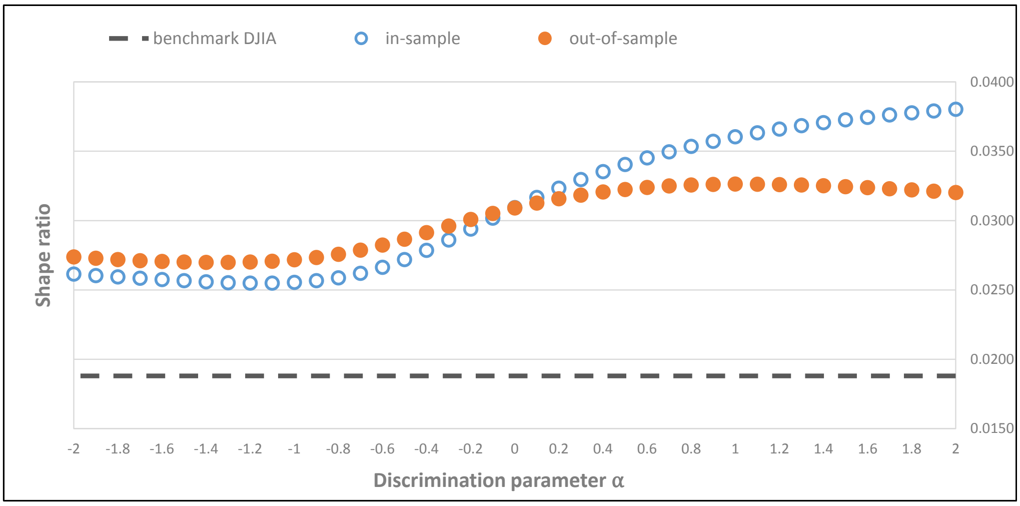

Figure 11: **Sharpe ratios for combined strategy.** The figure displays the behavior of the Sharpe ratio for portfolios constructed by utilizing Google and Wikipedia searches together in portfolio optimization. Both in-sample (empty blue circles) and out-of-sample (full orange circles) performances are shown in the figure. The discrimination parameter α ranges between -2 and 2with a step of 0.1. For $\alpha = 0$, the uniformly weighted portfolio is obtained. Black dashed line represents the performance of the DJIA strategy. Sharpe ratio maximizing portfolios are located at $\alpha = 2$ (in-sample) and $\alpha = 1$ (out-of-sample).

profit of 77%. As in the previous cases, both the in-sample and the out-of-sample approaches reach higher cumulative profit than the DJIA strategy, which yields the cumulative profit of 46%. Nonetheless, the difference between the out-of-sample performance and the DJIA strategy is important for drawing conclusions about the profitability of search-based strategies. Therefore, it should be noted that the difference in cumulative profits of 31% over more than 7 years seems not to be high enough to cover the costs connected with pursuing the search-based strategy. However, our strategies had not been constructed to maximize the profit, but rather to provide any utility to the portfolio diversification.

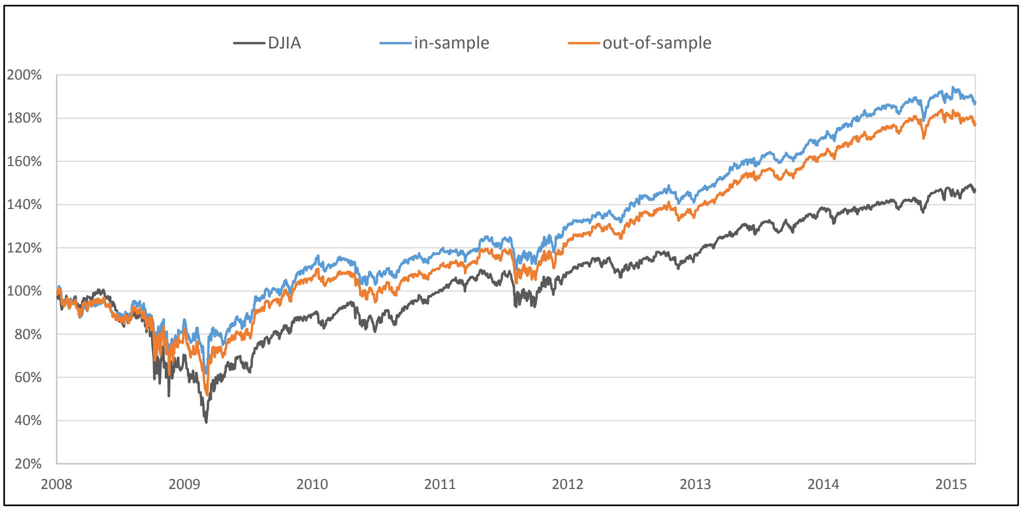

Figure 12: **Evolution of portfolios based on Google and Wikipedia searches.** Black line represents the evolution of the DJIA index, blue and orange lines show the development of the in-sample and out-of-sample strategies, respectively. Evolution is shown for the values of α which maximizes the Sharpe ratio for both the search-based strategies. For practical purposes, the comparison between the out-of-sample and the DJIA strategy is essential. The out-of-sample yields higher cumulative profit than the passive buy-and-hold strategy by about 30%.

6 Conclusion

In this thesis, we investigate whether the web search queries provided by Google Trends and Wikipedia are able to bring any advantages to portfolio diversification, a core technique used for minimizing portfolio risk. Using a sample of twenty-two DJIA companies, we employ a diversification rule that assigns weights to the stocks based on the popularity of stock-related queries on Google and Wikipedia. The idea behind is that the increase in searches for company-related keywords usually reflects that something is going on with the company, while this event will most probably affect the corresponding stock price. Therefore, more searched stocks are also suspected to be more volatile, hence discriminating them in the final portfolio should reduce the risk of the portfolio.

First, we measure the popularity of a stock by Google search volume for tickers; then we utilize tickers in combination with the word “stock”. Our results indicate that the search-based portfolios are able to reach lower risk level and higher standardized average returns than both the uniformly weighted portfolio and the benchmark DJIA index. Importantly, the diversification strategies were successful even in the out-of-sample. Furthermore, we demonstrate that discrimination of popular stocks pays off in the portfolio selection. This evidence essentially confirms the findings of Kristoufek (2013b) also at a daily timescale.

Second, we examine strategies based on Wikipedia pageviews of company-related articles. We show that portfolios diversified using Wikipedia pageviews dominate the benchmark DJIA and the uniformly distributed portfolio, both in-sample and out-of-sample. However, we observe the minimum variance portfolio for strategies preferring Wiki-popular stocks. Hence, we can conclude that Wikipedia pageviews proved to be worthy in portfolio optimization, but we cannot state that strategies based on the discrimination of Wiki-popular stocks are the best performing strategies. Considering the question which stocks (popular or peripheral) should be discriminated, the results are rather inconclusive. This can be explained by the fact that people usually visit Wikipedia to obtain some general knowledge about the company and not to find financial news. Accordingly, data on Wikipedia pageviews for company's page will probably reflect the information demand from a much broader audience than data on Google searches for tickers show. Nonetheless, the evidence supports the utility of web search data in portfolio diversification.

Lastly, we construct a combined measure of attention, which merges Google searches for tickers along with Wikipedia pageviews of company-related articles into one selection criterion (portfolio weight) according to which the portfolios are rebalanced. We again find portfolios that outperform the uniformly weighted portfolio and the benchmark DJIA index, both in-sample and out-of-sample. This further supports the hypothesis that search data can be successfully utilized for portfolio diversification.

For future research we leave examining if the ideal portfolio weights do change over time as well as empirical tests of the relationship between Wikipedia Trends data and stock price volatility. The specification we used can be easily extended to other stock market indexes (e.g. S&P 500) or different types of portfolios - portfolios including growing companies or companies from the less developed countries - if the data are available. Moreover, one can focus on other web search data like those from Yahoo! or Chinese search engine Baidu.

Bibliography

Alanyali, M., Moat, H.S., Preis, T., 2013. Quantifying the Relationship Between Financial News and the Stock Market. Scientific Reports 3.

Ap Gwilym, O., Hasan, I., Xie, R., Wang, Q., 2012. In search of concepts: The effects of speculative demand on returns and volume. Available at SSRN 2062813.

Askitas, N., Zimmermann, K. F., 2009. Google econometrics and unemployment forecasting. Applied Economics Quarterly 55 (2), 107-120.

Baker, S., Fradkin, A., 2011. What Drives Job Search? Evidence from Google Search Data. Discussion Papers 10-020, Stanford Institute for Economic Policy Research.

Bank, M., Larch, M., Peter, G., 2011. Google search volume and its influence on liquidity and returns of German stocks. Financial Markets and Portfolio Management 25 (3), 239-264.

Barber, B. M., Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. Review of Financial Studies 21 (2), 785-818.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., Weber, I., 2012. Web search queries can predict stock market volumes. PLoS ONE 7 (7), e40014.

Brownstein, J.S., Freifeld, C.C., Madoff, L.C., 2009. Digital Disease Detection - Harnessing the Web for Public Health Surveillance. New England Journal of Medicine 360 (21), 2153–2157.

Carneiro, H., Mylonakis, E., 2009. Google Trends: A web based tool for real time surveillance of disease outbreaks. Clinical Infectious Diseases 49 (10), 1557-1564.

Carrière-Swallow, Y., Labbé, F., 2010. Nowcasting with Google Trends in an emerging market. Central Bank of Chile Working Papers. Working Paper No. 588.

Chen, S., 2011. Google Search Volume: Influence and Indication for the Dutch Stock Market. Erasmus University.

Choi, H., Varian, H., 2009a. Predicting initial claims for unemployment benefits. Google Inc.

Choi, H., Varian, H., 2009b. Predicting the present with Google Trends. Google Research Blog 2.

Cooper, C. P., Mallon, K. P., Leadbetter, S., Pollack, L. A., Peipins, L. A., 2005. Cancer internet search activity on a major search engine, United States 2001-2003. Journal of Medical Internet Research 7 (3), e36.

Corley, C., Mikler, A. R., Singh, K. P., Cook, D. J., 2009. Monitoring influenza trends through mining social media. Proceedings of the 2009 international conference on bioinformatics and computational biology (BIOCOMP09); Las Vegas. NV.

Curme, C., Preis, T., Stanley, H. E., Moat, H. S., 2014. PNAS 111 (32), 11600–11605.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. The Journal of Finance 66 (5), 1461-1499.

Della Penna, N., Huang, H., 2009. Constructing a consumer confidence index for the US using web search volume. Tech. rep., Working Paper.

Dimpfl, T., Jank, S., 2011. Can internet search queries help to predict stock market volatility? Tech. rep., CFR working paper.

Dugas, A. F., Hsieh, Y.-H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., Gaydos, C. A., Perl, T. M., Rothman, R. E., 2012. Google flu trends: Correlation with emergency department influenza rates and crowding metrics. Clinical Infectious Diseases 54 (4), 463-469.

D'Amuri, F., Marcucci, J., 2010. Google it! Forecasting the US unemployment rate with a Google Job Search Index. Available at SSRN 1594132.

Ettredge, M., Gerdes, J., Karuga, G., 2005. Using web-based search data to predict macroeconomic statistics. Communications of the ACM 48 (11), 87-92.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012-1014.

Guzman, G., 2011. Internet search behavior as an economic forecasting tool: The case of inflation expectations. Journal of Economic and Social Measurement 36 (3), 119-167.

Hulth, A., Rydevik, G., Linde, A., 2009. Web queries as a source for syndromic surveillance. PloS one 4 (2), e4378.

Joseph, K., Wintoki, M. B., Zhang, Z., 2011. Forecasting abnormal stock returns and trading volume using investor sentiment: evidence from online search. International Journal of Forecasting 27 (4), 1116-1127.

Kristoufek, L. 2013a. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific Reports 3.

Kristoufek, L., 2013b. Can Google Trends search queries contribute to risk diversi- fication? Scientific Reports 3.

McLaren, N., Shanbhogue, R., 2011. Using internet search data as economic indicators. Bank of England Quarterly Bulletin (2011), Q2.

Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., Preis, T., 2013. Quantifying Wikipedia usage patterns before stock market moves. Scientific Reports 3.

Mondria, J., Wu, T., Zhang, Y., 2010. The determinants of international investment and attention allocation: Using internet search query data. Journal of International Economics 82 (1), 85-95.

Pavlicek, J., Kristoufek, L., 2015. Nowcasting unemployment rates with Google searches: Evidence from the Visegrad Group countries. FinMap – Working Paper No. 34.

Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., Valleron, A., 2009. More diseases tracked by using Google Trends. Emerging Infectious Diseases 15 (8), 1327-1328.

Peng, L., Xiong, W., 2005. Investor attention, overconfidence and category learning. Available at SSRN 724362.

Polgreen, P., Chen, Y., Pennock, D., Nelson, F., 2008. Using internet searches for influenza surveillance. Clinical Infectious Diseases 47 (11), 1443-1448.

Preis, T., Paul, W., Schneider, J. J., 2008. Fluctuation patterns in high-frequency financial asset returns. EPL (Europhysics Letters) 82 (6), 68005.

Preis, T., Reith, D., Stanley, H. E., 2010. Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368 (1933), 5707-5719.

Preis, T., Moat, H.S., Stanley, H.E., 2013. Quantifying trading behavior in financial markets using Google Trends. Scientific Reports 3.

Preis, T., Moat, H.S., 2014. Adaptive nowcasting of influenza outbreaks using Google searches. R. Soc. open sci. 140095.

Ramos, S. B., Veiga, H., Latoeiro, P., 2013. Predictability of stock market activity using google search queries. Tech. rep.

Shi, R., Xu, Z., Chen, Z., Huang, J., 2012. Does attention affect individual investors' investment return? China Finance Review International 2 (2), 143-162.

Smith, G. P., 2012. Google internet search activity and volatility prediction in the market for foreign currency. Finance Research Letters 9 (2), 103-110.

Suhoy, T., 2009. Query indices and a 2008 downturn: Israeli data. Research Department, Bank of Israel.

Vakrman, Tomáš. Google searches and financial markets: IPOs and uncertainty. Prague, 2014. 173 pp. Master Thesis (Mgr.). Charles University in Prague, Faculty of Social Sciences, Institute of Economic Studies. Supervisor: PhDr. Ladislav Krištoufek Ph.D.

Vakrman, T., Kristoufek, L., 2015. Underpricing, underperformance and overreaction in initial public offerings: Evidence from investor attention using online searches. SpringerPlus 4.

Vlastakis, N., Markellos, R. N., 2012. Information demand and stock market volatility. Journal of Banking & Finance 36 (6), 1808-1821.

Vosen, S., Schmidt, T., 2011. Forecasting private consumption: survey-based indicators vs. Google Trends. Journal of Forecasting 30 (6), 565-578.

Wilson, B.J., 2009. Early detection of disease outbreaks using the internet. Canadian Medical Association Journal 180 (8), 829-831.

Zhang, W., Shen, D., Zhang, Y., Xiong, X., 2013. Open source information, investor attention, and asset pricing. Economic Modelling 33, 613-619.

Zhou, X., Ye, J., Feng, Y., 2011. Tuberculosis surveillance by analyzing Google Trends. IEEE transactions on bio-medical engineering.

Appendix A

Companies in DJIA sample and their characteristics

Table 1: **Companies in the DJIA sample, their tickers and corresponding Wikipedia page titles.** List of companies in the DJIA sample (1), their corresponding tickers (2) and titles of Wikipedia pages (3).

Company (1)	Ticker (2)	Wikipedia's page title (3)
American Express Company	AXP	American Express
The Boeing Company	BA	Boeing
Caterpillar Inc.	CAT	Caterpillar Inc.
E. I. du Pont de Nemours and Company	DD	DuPont
The Walt Disney Company	DIS	The Walt Disney Company
General Electric Company	GE	General Electric
The Home Depot, Inc.	HD	The Home Depot
International Business Machines Corporation	IBM	IBM
Intel Corporation	INTC	Intel
Johnson & Johnson	JNJ	Johnson & Johnson
JPMorgan Chase & Co.	JPM	JPMorgan Chase
The Coca-Cola Company	KO	The Coca-Cola Company
McDonald's Corp.	MCD	McDonald's
3M Company	MMM	3M
Merck & Co. Inc.	MRK	Merck & Co.
Microsoft Corporation	MSFT	Microsoft
Pfizer Inc.	PFE	Pfizer
The Procter & Gamble Company	PG	Procter & Gamble
United Technologies Corporation	UTX	United Technologies Corporation
Verizon Communications Inc.	VZ	Verizon Communications
Wal-Mart Stores Inc.	WMT	Walmart
Exxon Mobil Corporation	XOM	ExxonMobil

Appendix B

Thesis Proposal

Thesis Proposal

Kristýna Brunová

6.6.2014

Proposed topic

Are the more popular stocks also the more risky ones? Google and Wikipedia searches in portfolio optimization

Supervisor

PhDr. Ladislav Krištoufek Ph.D.

Topic characteristics

In the most recent years, social sciences have obtained access to huge data sets based on internet activity of millions of users all over the world. These massive new data sources can offer a new perspective on the behavior of people. In this thesis we will use the information from Google Trends, a service that shows the popularity of searched terms, and Wikipedia to construct a portfolio diversification strategy. We will propose a portfolio diversification strategy based on the search volume of stock-related terms. The diversification strategy stems in an idea that the more frequently the stock-related term is searched for the higher the risk (in the financial perspective) of a specific stock. According to this assumption, that the popular stocks should be discriminated in the final portfolio, we will assign them lower weights to decrease the total risk of the portfolio. Finally the comparison between search-based portfolios and the constant buy-and-hold strategy will be discussed. The aim of this thesis is to answer the question how can we use data from Google Trends and Wikipedia in portfolio optimization? Can search queries be utilized in portfolio selection and risk diversification?

Hypotheses

1. How does the popularity of the stock contribute to its riskiness?
2. Is the search-based strategy of portfolio diversification more valuable than a standard buy-and-hold strategy?
3. Does the strategy outperform a uniformly distributed portfolio?

Methodology

We compute portfolio weights based on Google and Wikipedia search volumes for stock-related terms. Regarding Google, we use the ticker symbol of a stock and the combination of the word "stock" and the ticker symbol. Then we utilize Wikipedia pageviews of company-related pages. According to these weights, portfolio are rebalanced. Subsequently, we compare the performance of the search-based portfolios with the passive buy-and-hold strategy and the uniformly distributed portfolio. We will try to analyze as large data sets as possible to confirm our hypothesis.

Core bibliography

Kristoufek, L., 2013. Can Google Trends search queries contribute to risk diversification? Scientific Reports 3.

Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., Preis, T., 2013. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. Scientific Reports 3.

Preis, T., Reith, D., Stanley, H. E., 2010. Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368 (1933), 5707-5719.

Dimpfl, T., Jank, S., 2011. Can internet search queries help to predict stock market volatility? Tech. rep., CFR working paper.