IDEALIZE - A NOTION OF IDEA STRENGTH

Rui Portocarrero Sarmento
LIAAD-INESC TEC
PRODEI - Faculty of Engineering, University of Porto
mail@ruisarmento.com

April 9, 2019

ABSTRACT

Business Entrepreneurs frequently thrive on looking for ways to test business ideas, without giving too much information. Recent techniques in startup development promote the use of surveys to measure the potential client's interest. In this preliminary report, we describe the concept behind Idealize, a Shiny R application to measure the local trend strength of a potential idea. Additionally, the system might provide a relative distance to the capital city of the country. The tests were made for the United States of America, i.e., made available regarding native English language. This report shows some of the tests results with this system.

Keywords Business · Trends and Keywords · Entrepreneurship

1 Introduction

Business Idea strength is a function of several variables. One of these variables, the local strength, is a major variable when entrepreneurs are dealing with establishing a business. Some businesses vary the weight they consider this variable to have. For example, retail might give more weight to this than services.

With the advent of internet search at the beginning of the XXI century, clients rely more and more on the Google search engine. This service provides locally oriented searches, and provide a list of available businesses that might fulfill the potential client needs.

In this report we test several assumptions, based on Google search engine and an index of local keyword strength, given an input summary for a business idea.

This document is organized in the following manner: Section 2 presents a small summary of previous work in this area. Then, in section 3, we present the tools we gathered to develop the testing system. Section 4 presents some of the features of the developed system. Section 5 gives an idea of what results to expect from the system. Finally, in section 6, we discuss the main advantages and drawbacks of the solutions and give some suggestions for future work. Section 7 concludes this document.

2 Background

Some works have been done in the attempt to use some measures of search to improve prediction of outcome in several areas.

From Quora (2019), we can read that some researchers stress that Google Trends data needs to be interpreted in the context in which a keyword is used, as well as the overall context of the research carried out.

In the context of financial research, some authors have suggested that Google Trends data can be interpreted as a description of collective behavior, aggregate demand, stock market moves, investor expectations, information demand, attention, or market sentiment (Curme et al., 2014; Carrière-Swallow and Labbé, 2010; Preis et al., 2013; Siganos, 2013; Schmidt and Vosen, 2009; Vlastakis and Markellos, 2012; DRAKE et al., 2012; Choi and Varian, 2012).

Some of the quoted authors in Quora (2019) refer, as examples:

- "Internet search data may offer new possibilities to improve forecasts of collective behavior"
- "Our findings /.../ suggests that Google data is a promising source of information for nowcasting components of aggregate demand in short-run models"
- "By analyzing changes in Google query volumes for search terms related to finance, we find patterns that may be interpreted as 'early warning signs' of stock market moves."
- "We use a novelty Google search volume to proxy the market expectation hypothesis according to which firms with an abnormal upward change in Google searches are identified as firms with potential merger activity."
- "[W]e introduce a new indicator for private consumption based on search query time series provided by Google Trends."
- "Demand is approximated in a novel manner from weekly internet search volume time series drawn from the recently released Google Trends database."
- "We propose a new and direct measure of investor attention using search frequency in Google"
- "We use daily internet search volume from millions of households to reveal market-level sentiment."
- "How to use search engine data to forecast near-term values of economic indicators. Examples include automobile sales, unemployment claims, travel destination planning, and consumer confidence."

3 Tools

In this section, we introduce some of the techniques we used in our tests. From software to concepts explored, everything is cited, and the methodology is described here.

3.1 TextRank

Figure 1. Original TextRank workflow

In Mihalcea and Tarau (2004), the authors present the developed TextRank, as a solution to obtain keywords automatically from text. Figure 1 shows the workflow of the original TextRank algorithm. The text corpus processing starts by the pre-processing of the text and the removal of stop words, numbers and punctuation. Then, the document goes through a process of annotation where remaining single words are categorized, for example, like nouns, verbs or adjectives among others. This method is called Part-of-Speech tagging (POS tagging). According to the authors, only a few of these annotated words are essential. The authors studied which group of words delivered the best results, and they concluded that the best automatic keyphrases were obtained with nouns and adjectives. Then, with these filtered words, a graph-based approach is used. Each word is considered a graph node and the connections of words in this directed graph is determined by the order they appear in the text. The weight of these links is obtained by counting the number of times these pairs of words occur in the text corpus. The next phase of the algorithm regards the selection of the words of high importance. This is done with the use of the PageRank algorithm by Page et al. (1998). The words with high PageRank values are selected as potential keywords. Finally, the keyphrases are obtained with a post-processing stage. This stage involves the use of a sliding window evolving through the initial text to assess the order of words that are contained in the keyphrases or keywords. This step takes into account punctuation and other structural features of the document to retrieve reasonable keyphrases. TextRank has been widely praised as a consistent method to automatically retrieve keywords from the text. Inclusively, it has been used in prototypes of decision support systems as narrated by Brazdil et al. (2015).

3.2 Google Trends

We used an R package (gtrendsR) for retrieving Google trends regarding input keywords. This makes a potentially good notion of the trend, in search hits, a particular keyword has. For more information, please see Philippe Massicotte (2018).

Interest over time, a measure retrieved from Google Trends, is calculated as follows:

$$\text{Interest}(Keyword)_t = \frac{\#of\,queries\,for\,keyword_t}{TotalGoogleSearchQueries} \tag{1}$$

Search interest is both indexed and normalized. This means the particular interest at time t, is divided by the maximum number of interest in the interval of search x. Equation 1 becomes equation 2, and this is how Google retrieves results of interest over time for a particular keyword.

$$\text{NormalizedInterest}(Keyword)_{t\in[t-x,t[} = \frac{\#of\,queries\,for\,keyword_{t\in[t-x,t[}}{\max \#of\,queries\,for\,keyword_{[t-x,t[}} * 100 \tag{2}$$

For this reason, with regional analyses, we are retrieving a normalized indication of search interest within that particular country. An interesting index of 100 in Portugal and an index of 55 in Spain, would mean that the concentration of Portuguese searching for a particular keyword is higher than the concentration of Spanish searching for the same keyword. It can mean that Spanish are less interested in the keyword, or they may search for way more other concepts. The measure doesn't take into account the difference between countries in the size of the internet population and volume of queries per user.

3.2.1 Normalized Keyword Weight

Since the TextRank algorithm retrieves several keywords or keyphrases, we can achieve a weight we can use in the calculus of average trend; we will approach this in next subsections. Therefore, we use a Normalized Keyword Weight (NKW), from the TextRank algorithm:

$$\text{NKW}_{\text{keyword}_k} = \frac{Weight_{keyword_k}}{\sum_{k=1}^{nkeywords} Weight_{keyword_k}} \tag{3}$$

3.2.2 Average Trend per Idea

We achieve an average trend of our idea by adding weighted trend values per keyword, in the interval selected for search. The weight is retrieved from the TextRank algorithm, that gives us the weighted keywords or Keyphrases, given by TextRank weights. Thus,

$$\text{AverageTrendPerIdea}(Idea)_{t_n} = \frac{\sum_{k=1}^{nkeywords} NKW_{keyword_k} * NormalizedInterest(keyword_k)_{t_n}}{nkeywords} \tag{4}$$

4 Developed System

In figure 2, we can see the Idealize application [1]. There are two main regions in the application interface, one on the left of the user and another to the right of the screen.

On the left region (figures 3,4 and 5), the user can select a country or the world, regarding region of study. The user can also select the context of a search for the keywords extracted from the text, and finally, the user can select the timeframe for the searches.

Regarding Context Selection, the user can select one of the following choices:

- "web"
- "news"
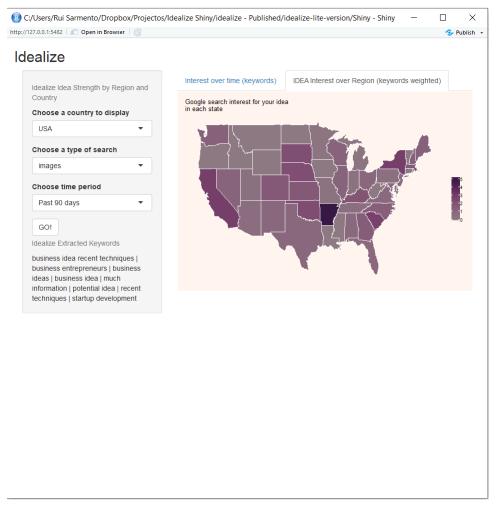- "images"
- "froogle"
- "youtube"

Figure 2. Idealize Application

Regarding Timeframe Selection, the user can select one of the following choices:

- "Last hour"

- "Last four hours"

- "Last day"

- "Last seven days"

- "Past 30 days"

- "Past 90 days"

- "Past 12 months"

- "Last five years"

- "Since the beginning of Google Trends (2004)"

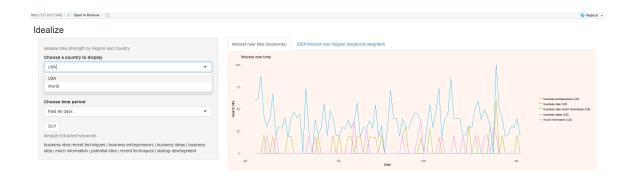[1]Available Code at `https://github.com/Sarmentor/idealize-lite-version`.
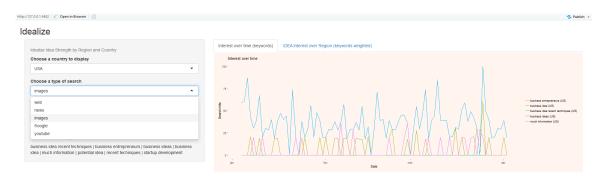
Figure 3. Idealize Application - Country Selection

Figure 4. Idealize Application - Search Context Selection

5 Results

5.1 Idea Examples

As an example, the input text was changed with two tentative concepts/ideas. We started by inputting this report's abstract. Then we inputted another conceptual idea, this time regarding automobile services. From the keywords in the texts, we extracted plots of keyword trends, and also the plots for the idea strength per region of study.

5.1.1 Input Text I

"Business Entrepreneurs frequently thrive on looking for ways to test business ideas, without giving too much information about their business idea. Recent techniques in startup development promote the use of surveys to measure the potential client's interest in a business idea. In this preliminary report, we describe the concept behind Idealize, a Shiny R application to measure the local trend strength of a potential idea for a business. Additionally, the system might provide a relative distance to the capital city of the country. The tests were made for the United States of America, i.e., made available regarding native English language. This report shows some of the tests results with this system."

Trends and maps for the USA

In Figures 6 and 7 (see APPENDIX), we show the results for this text, regarding a 5 year timeframe and for web searches.

Figure 6 shows the trend chart of several extracted keywords; it is relatively clear that some keywords have different trends, some more seasonal than others. It is also visible we show only five keywords. This is due to a limitation in the gtrendsR package that allows a limit of 5 keywords searches at a time.

Figure 7 shows the idea strength map. It is shown that some states give more importance to such an idea than others. As darker the color in the selected color gradient, the more importance is given to the user idea.
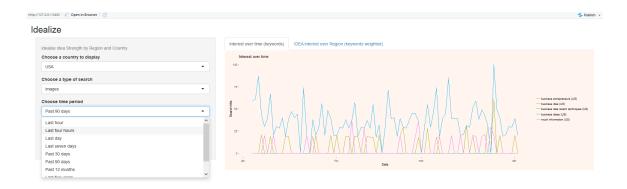
Figure 5. Idealize Application - Timeframe Selection

In Figures 8 and 9 (see APPENDIX), we show the results for this same text, this time regarding a 1-year timeframe and Youtube searches.

Figure 8 shows the trend chart of the several extracted keywords. As expected, keywords have different trends, again, some more seasonal than others.

Figure 9 shows the idea strength map. It is shown the importance that states give to such an idea.

5.1.2 Input Text II

"Our company provides the best service in auto maintenance and tunning. We have several stores around the country. We serve our clients all auto brands, either regarding car parts or maintenance.

In the tunning department, we provide several tunning brands and also provide tunning services as per the client needs."

Trends and maps for the USA

In Figures 10 and 11 (see APPENDIX), we show the results for this text, regarding a 1-year timeframe and web searches.

Figure 10 shows the trend chart of the several extracted keywords; it is relatively clear that some keywords have different trends, comparatively with the first idea results.

Figure 11 shows the second idea strength map. Comparatively, with the first idea, we can see that with the idea concept change, several changes were produced in the distribution, per state, regarding idea strength. This is the expected result, with the use of this system.

6 Discussion

This report brings some light to some possible methods to approach idea development from its inception. Although several other approaches are possible, this one involves very low effort. A logical synopsis or summary of an idea and a location is all that is needed. Nonetheless, as reported, this system has several concepts that are far from perfect; for example, there are some base measures that are simple approaches and not exact. Moreover, some values change subjectively, with variations of text input and English language grammar. This is true, although the base conceptual idea to be tested might be the same, but explained with different words.

Additionally, regarding the evaluation of the model, we did not have the time or resources available to retrieve ground truth data to test our assumptions. This would be possible, for example, by accessing startups or SME's temporal data of results. This would be important to fine tune some tasks in the several phases of the workflow in the proposed system.

7 Conclusions

This academic report, about the experiment with idea strength measurement, brought light to some methods for achieving a notion about an idea of business. This proposed low-cost method is interesting but exposes many issues or doubts about its value. There are many variables that might be added in the future to such a system. No system is perfect, and the complexity of such a prediction as idea strength has many factors involved. The presented solution does not entitle itself as the final solution, as discussed in this report, many other factors might have to be used to achieve a good prediction of idea strength.

Acknowledgments

References

Brazdil, P., Trigo, L., Cordeiro, J., Sarmento, R., and Valizadeh, M. (2015). Affinity mining of documents sets via network analysis, keywords and summaries. *Oslo Studies in Language*, 7(1).

Carrière-Swallow, Y. and Labbé, F. (2010). Nowcasting with google trends in an emerging market. Working papers central bank of chile, Central Bank of Chile.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *The Economic Record*, 88(s1):2–9.

Curme, C., Preis, T., Stanley, H. E., and Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, 111(32):11600–11605.

DRAKE, M. S., ROULSTONE, D. T., and THORNOCK, J. R. (2012). Investor information demand: Evidence from google searches around earnings announcements. *Journal of Accounting Research*, 50(4):1001–1040.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia.

Philippe Massicotte, D. E. (2018). *gtrendsR: Perform and Display Google Trends Queries*. R package version 1.4.2.

Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Sci. Rep.*, 3.

Quora (2019). How do you interpret Google Trends's search volume index? - Quora. [Online; accessed 16. Mar. 2019].

Schmidt, T. and Vosen, S. (2009). Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. Ruhr Economic Papers 155, RWI - Leibniz-Institut für Wirtschaftsforschung, Ruhr-University Bochum, TU Dortmund University, University of Duisburg-Essen.

Siganos, A. (2013). Google attention and target price run ups. *International Review of Financial Analysis*, 29:219–226.

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

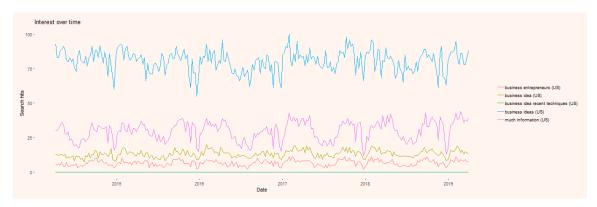APPENDIX

A RESULTS FIGURES

A.1 IDEA I

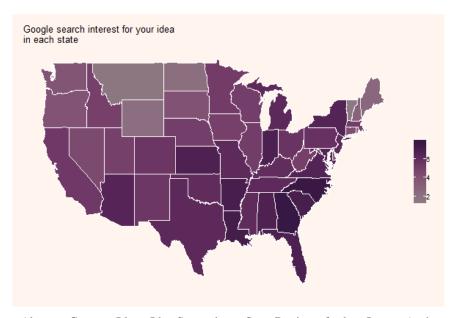**Figure 6.** Abstract Concept Idea - Keyword Trends - for last 5 years (web searches)

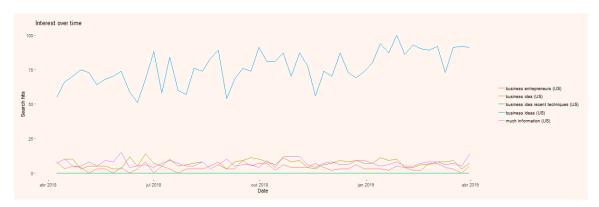**Figure 7.** Abstract Concept Idea - Idea Strength per State Region - for last 5 years (web searches)

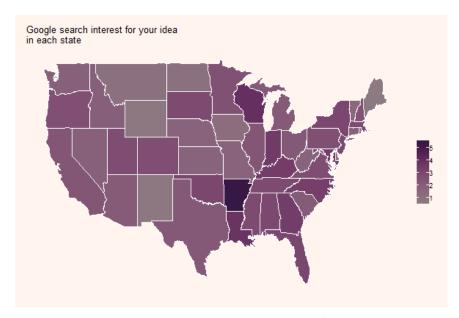**Figure 8.** Abstract Concept Idea - Keyword Trends - for last 1 year (Youtube searches)

Figure 9. Abstract Concept Idea - Idea Strength per State Region - for last 1 year (Youtube searches)
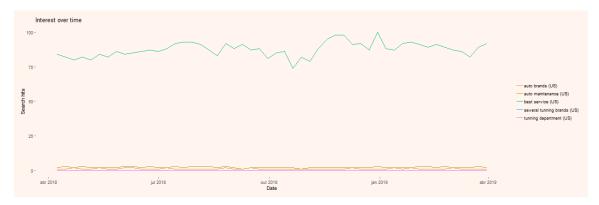
A.2 IDEA II

Figure 10. Auto Concept Idea - Keyword Trends - for last 1 year (web searches)
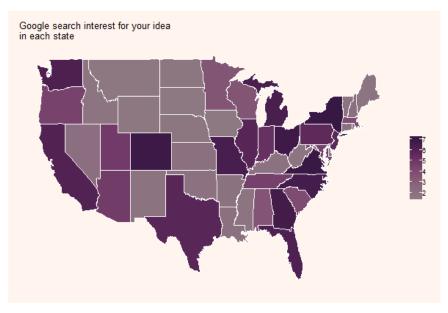
Figure 11. Auto Concept Idea - Idea Strength per State Region - for last 1 year (web searches)