Identifying text phrases containing similar semantic meaning with search keywords using NLP

Y. T. L Somarathna

Index No : 16001402

Supervisor: Dr M. G. N. A. S. Fernando

<February 2021>

Submitted in partial fulfillment of the requirements of the

B.Sc in Computer Science Final Year Project (SCS4124)

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: Y. T. L. Somarathna

……………………………………………….

Signature of Candidate Date:

This is to certify that this dissertation is based on the work of

Dr. M. G. N. A. S. Fernando

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Principle/Co- Supervisor's Name: Mr W.V Welgama

……………………………………………….

Signature of Supervisor Date:

Abstract

A core strategy of content SEO is placing the right amount of keywords in a webpage to archive the optimal keyword density with related quality keywords. The current approach for performing this task is manually going through the content and checking for text phrases that can be replaced by keywords and replacing them. In this study, we focus on a specific domain of websites (Android) and try to develop a model that can automatically recognize text phrases that are semantically similar to keywords. Distributional semantics is used to capture the meaning of words and a un-supervised classifier is used to classify phrases related to a particular keyword with the help of words semantics captured by distributional semantics.

Even with the latest NLP advancement capturing and comparing the semantics of text phrases seems to be some problems as text phrases have less context to work with. This study shows that still, we can produce useful results in recognizing text phrases with similar semantic meaning with search keywords.

Preface

This dissertation has been written for the partial fulfilment of the requirements of the B.Sc. in Computer Science (Hons) Final Year Project in Computer Science (SCS4124). I was engaged in this research and writing this dissertation from April 2020 to February 2021.

This research work presents the studies of applying and deep learning techniques to identify word phrases which share a similar meaning with search keywords. To best of my knowledge, research work on automating the keyword process of Search Engine Optimization using deep learning approach or any other approach has not been carried out so far. Therefore, this work will be the first of it's kind. We have created our dataset containing a large number of pairs of keywords and word phrases with similar meaning in the android domain for this study. We created a model using existing deep neural architecture BERT for classifying word phrases to the relevant keywords. Further, we did several experiments with this model and applied various steps to fine-tune the models to enhance the output. The result presented in Section 5 relies upon experiments conducted by me.

With constant guidance and supervision of my supervisor and co-supervisor, conclusions were drawn on evaluation and training the models. This piece of research work would be a great source of knowledge for future research on using NLP in SEO.

Acknowledgement

First, I would like to thank my university, University of Colombo School of Computing (UCSC) for giving me this great opportunity to carry out individual research in which I could develop my research and other academic skills.

I would like to express my sincere gratitude to my research supervisor Dr M. G. N. A. S. Fernando and my co-supervisor Mr W.V Welgama for providing me with their valuable guidance and supervision throughout this research project. I am grateful to them for finding time to help me all the time, from the beginning to the end of this research project.

Further, I would like to pay my sincere gratitude to Dr. H.N.D. Thilini, computer science project coordinator, and all other UCSC staffs members for all the assistance they provided to make this project successful.

Table of Contents

List of Figures

List of Tables

List of Acronyms

Chapter 1 - Introduction

1.1 Background to the Research

The web is changing rapidly both the content and the technologies used, So does the users and the search engines, which facilitates the users. Search engines use complex algorithms to determine the relevance of a certain document related to a particular search query and displays the relevant documents ranked by their importance. Those algorithms consider several factors to determine the importance of each document such as the keywords used within the content and number of backlinks a particular website has. And those factors are constantly changing and evolving throughout the last decade. And the webmasters(Web Content Creators) are using various Search Engine Optimization(SEO) techniques to evolve their content to meet new standards of search engines to reach their audience. Studies have been carried out to determine the importance of SEO as a whole and also to determine the importance of various SEO strategies.

In [1] states that the ultimate SEO goal is to provide the basic policy to optimize websites, in order for the latter to succeed in higher and better-related rankings in the search engines, as well as better targeted traffic, both in volume and depth.

The ratio between words in all the keywords of a webpage and all the words within the webpage is called the keyword density of a web page. Properly placing keywords within the content to archive the optimal keyword density and using the right set of keywords in that process is an essential part of any webmasters SEO strategy. As the current approach which is done manually by an expert in SEO by analyzing keywords related to the niche of the website, and changing the content accordingly.

With the advancements of NLP, this study is focused on the first and major step of automating this process, which is building a system which can reliably identify text phrases which share similar semantic meaning with a search keyword.

1.2 Research Problem and Research Questions

Becoming skilled at SEO or hiring an expert on SEO has become an essential thing for any content creator or marketer on the internet to reach potential organic viewers, which can be very costly and time-consuming. So having a tool which can automatically optimize content for its relevant search metrics can benefit both the content creator and the reader at the same time as saves time and cost for SEO optimizations and also it brings more relevant content for the reader. This study does not address the whole SEO optimization process.

But containing the proper keyword cluster in the optimal density within the web pages is the main limiting factor of the content creation as it takes a lot of time and effort to complete. Even though there are many other important aspects of SEO such as backlink creation and implementing search engine friendly architecture within the website, all those major SEO practices can be separated from the content creation and executed independently.

This research project intends to analyze the applicability of semantic matching based on distributional semantic NLP techniques to identify word phrases which can be replaced by word phrases containing keywords. As the first step to automate the process of optimizing the keyword density of web pages.

To address the requirement above, the proposed study will answer the following questions

- How to identify word phrases that contain a similar meaning with keywords related to content?
  - To identify a text phrase which shares a meaning with a search keyword we must have a way of measuring the semantic similarity between those two test phases.

- What are the techniques available to compare the similarity of texts?

    - Measuring the semantic similarity of texts(semantic matching) is categorized under NLP and has a lot of approaches for archiving this, which are based on various underlying techniques which are used appropriately as they suite the desired end result going to be archived by the semantic matching task.

- Which of those are applicable to compare semantic similarity between small word phrases?
    - Most of the semantic matching techniques are not suitable for short text similarity tasks especially the techniques which are not based on lexical matching, as they are not able to capture the semantic similarity over a trivial level.
- How to develop a model that can identify similarities between short text phrases and search keywords from those applicable techniques?
    - By combining the power of distributed semantics to capture the semantics of words and encoding them into text embeddings using an appropriate encoding technology and using an additional layer to classify text into relevant keywords.

1.3 Justification for the research

With the advancement of Artificial intelligence and machine learning, computers are becoming good at tasks which were seemingly impossible to perform by a computer. Which makes machine learning the go-to solution for most modern computer science problems.

This study explores the applicability of NLP techniques to enhance the keyword density of webpages and increase the overall SEO ranking of the webpages which reduces the time and effort needed to reach their full potential audience. Which saves them time and effort to create better and more content.

Human-error is a significant factor in SEO where final results highly depend on the expertise and the experience of the person who performs the SEO optimizations. So an accurate software solution which can perform this can omit that human-error.

Results of this study can enable and inspire studies on the inapplicability of NLP and machine learning in other areas of content SEO. Especially with the ability to understand the semantics of short word phrases related to the keywords, It can be used to filter out useful comments which actually add value for the content from spammy or unuseful comments, and also for automating the creation of internal links within the content.

And semantic matching over short text phrases seems to be attracting less attention throughout the literature. So the results of this study can also contribute to widening the understanding of semantic matching as it explores it in a not widely used context.

1.4 Methodology

As the first step a sufficient datasets for training the model and evaluating the model must be created. However expanding this study over every niche of websites is not practical with this study. So websites of the niche “Android” have been chosen to carry out the research because it meets the following challenges of the study can be met with the “Android” niche.

- Need a domain understanding in creating datasets
- Has a large amount of content and a large audience which consumes the content
- Vocabulary used in the field changes rapidly

A dataset of keywords about the Android niche is created using keywords extracted from top websites about Android using Adword Keyword Planner[2]. And those keywords will be combined with word phrases, with the help of domain knowledge about “Android” to manually create the a dataset consisting of keywords and word phrases with similar semantic meaning with keywords to evaluate the implemented model.

The proposed solution consists of two parts. Word embeddings which are obtained by unsupervised learning and a classifier to classify word phrases to the keyword it belongs to. Both keywords and phrases to be matched needed to be encoded to a mathematical representation of the text phrases using the most suited mechanism described in the literature, in order to be classified.

According to the literature review carried out, BERT model is one of the best achievements of modern deep learning-based NLP, which produce the state of the art solutions for a lot of problems of NLP has the best ability to capture the semantics. So I decided to use a pre-trained BERT model to capture the semantic of word phrases.

So In this study a pre-trained BERT model is used to capture the semantics of keywords and candidate word phrases to classify as share the semantic meaning with a particular keyword as vectors of semantics. Further details about this can be found in chapter 3. So with this semantic vectors, we can use a distance measure such as euclidian distance to match the best semantically matching keyword or top matching keywords for a given phrase. But here another problem arise, that is the number of keywords for a given niche is huge so holding the semantic vectors of the entire list of keywords is memory intensive and not practical, and embedding each keyword at the time of getting the distance measure is so time-consuming, so we have to optimize the problem on memory and time.

As a solution for this, a custom implemented k-means algorithm is used to cluster the keyword set into several centroids only the vectors of centroids were kept in the memory. So as the first step of the classification a winning centroid will be calculated and then all the keywords in that particular centroid will be encoded with the BERT model and best matching keywords can be found from here by applying a distance measure with those keywords semantic vectors.

1.5 Outline of the Dissertation

This thesis is organized as follows, In Chapter 2, a comprehensive study about existing techniques and approaches related to the domain of NLP and SEO is presented. The research design, along with the high-level architecture for addressing their search question, is presented in Chapter 3. Chapter 4 demonstrate the implementation details of the proposed methodology. In Chapter 5, comprehensive details of experiments carried out and evaluation result of all the model implemented is presented. Last chapter, Chapter 6 demonstrate the conclusion of the research and outlines future work.

1.6 Definitions

1.7 Delimitations of Scope

In Scope

- Analysing the applicability of currently available distributed semantic techniques on capturing semantic similarity between phrases to find phrases which can be replaceable by phrases with keywords.
- Building a machine learning model which is capable of enhancing the keyword density of web pages by identifying word phrases which are replaceable by keywords.
- Training the above-implemented model and evaluating the results.

Out of Scope

- This study is only applicable to content made in English and does not intend to support multilingual content.
- This study does not focus on the other aspects of SEO such as backlink building and neither on improving the content quality or user experience.

1.8 Conclusion

Chapter 2 - Literature Review

Chapter 3 - Design

Chapter 4 - Implementation

Chapter 5 - Results and Evaluation

Chapter 6 - Conclusions

6.1 Introduction

6.2 Conclusions about research questions (aims/objectives)

6.3 Conclusions about research problem

6.4 Limitations

6.5 Implications for further research

References

[1] T. Mavridis and A. L. Symeonidis, “Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms,” *Eng. Appl. Artif. Intell.*, vol. 41, pp. 75–91, May 2015, doi: 10.1016/j.engappai.2015.02.002.

[2] “Keyword Planner - 272-369-4930 - Google Ads.” https://ads.google.com/aw/keywordplanner/home?ocid=110535901&euid=115074181&__u=5045184669&uscid=110535901&__c=2656784949&authuser=0&subid=lk-en-ha-aw-bk-c-l00%21o3~Cj0KCQjw6PD3BRDPARIsAN8pHuGOdL8MdP1bVt5K_ZPA37DkJsBF9KJPvBlGBhtRZyjXKe8sTGmOHLgaAuFAEALw_wcB~76455040501~kwd-60278213695~6806576082~389349835136 (accessed Jul. 02, 2020).

Appendix A: Publications

Appendix B: Diagrams

Appendix C: Code Listings