The influence of search engine optimization on Google's results:

A multi-dimensional approach for detecting SEO

Dirk Lewandowski
Department of Information, Hochschule für Angewandte Wissenschaften Hamburg, Hamburg, Germany
dirk.lewandowski@haw-hamburg.de

Sebastian Sünkler
Department of Information, Hochschule für Angewandte Wissenschaften Hamburg, Hamburg, Germany
sebastian.suenkler@haw-hamburg.de

Nurce Yagci
Department of Information, Hochschule für Angewandte Wissenschaften Hamburg, Hamburg, Germany
nurce.yagci@haw-hamburg.de

ABSTRACT

Search engine optimization (SEO) can significantly influence what is shown on the result pages of commercial search engines. However, it is unclear what proportion of (top) results have actually been optimized. We developed a tool that uses a semi-automatic approach to detect, based on a given URL, whether SEO measures were taken. In this multi-dimensional approach, we analyze the HTML code from which we extract information on SEO and analytics tools. Further, we extract SEO indicators on the page level and the website level (e.g., page descriptions and loading time of a website). We amend this approach by using lists of manually classified websites and use machine learning methods to improve the classifier. An analysis based on three datasets with a total of 1,914 queries and 256,853 results shows that a large fraction of pages found in Google is at least probably optimized, which is in line with statements from SEO experts saying that it is tough to gain visibility in search engines without applying SEO techniques.

CCS CONCEPTS

• **Information systems** → World Wide Web; Web searching and information discovery; Content ranking; Information retrieval; Evaluation of retrieval results; World Wide Web; Web mining.

KEYWORDS

Search engines, search engine optimization, screen scraping

ACM Reference Format:
Dirk Lewandowski, Sebastian Sünkler, and Nurce Yagci. 2021. The influence of search engine optimization on Google's results:: A multi-dimensional approach for detecting SEO. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447535.3462479

1 INTRODUCTION

Search engine optimization (SEO), which is defined as "the practice of optimizing web pages in a way that improves their ranking in the organic search results" [26], can be regarded as lying somewhere between helping search engines finding and indexing relevant content and manipulating their results. The SEO industry's revenue is expected to reach $80 billion in the U.S. in 2020 [30]. Many websites heavily depend on the traffic they gain from search engines, especially from Google, the market leader in Web search. This search engine has a market share of 87% in the United States and 93% in Europe across all platforms as of January 2021 [49].

While there are no studies investigating the overall traffic websites receive from search engines, search traffic to some popular sites gives an indication: The New York Times website receives 33.5% of its traffic from search engines, IBM.com 45.6%, pinterest.com 38.0%, and acm.org even 67.6% [47].

A central question that has not been addressed in prior research is to which degree result lists in search engines are externally influenced by search engine optimization (SEO). While there is a vast body of professional literature on optimizing websites to be better found through search engines, the scholarly literature on SEO mainly focuses on analyzing search results to help website owners apply scientifically proven SEO techniques to improve their sites' rankings.

In our research, we take the user's perspective on SEO: To what degree are search engine result lists influenced by optimized pages? This is an essential question as the influence of SEO could mean that high-quality pages that have not been optimized are suppressed from the top results and replaced by lower-quality pages. This, in turn, would take into question the success of search engines in providing users with the most relevant results.

This paper aims to describe our approach to identifying content optimized for search engines through analyzing the HTML code and further data on the website level, combined with lists of manually classified websites. We further classify results by using a decision tree with features of relevant factors for SEO measures. Our analysis is based on three datasets containing a total of 256,853 URLs. The main results are that a large fraction of top results in Google has been optimized, that we can assume the effect of SEO to be stronger for popular queries, and that there are no vast differences between the ratio for optimized results on different result positions.

The rest of this paper is structured as follows: First, we provide a literature review showing the background of our research. Then, we report on how we identified and classified SEO indicators. After that, we give information on the URL classification and the datasets used. We then present the results of our analysis, discuss them and conclude with implications and suggestions for further research.

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom

ACM ISBN 978-1-4503-8330-1/21/06…$15.00
https://doi.org/10.1145/3447535.3462479

2 LITERATURE REVIEW

This section gives an overview of research relevant to understanding the practices and the influence of search engine optimization, user selection behavior on result pages, trust in search engines, and approaches used to identify SEO in search result documents.

2.1 The practice and influence of search engine optimization

In the context of search engine marketing (SEM), there are two ways to gain visibility on the result pages of commercial search engines: paid search marketing (PSM) and search engine optimization (SEO). In PSM, content provider book ads with the search engine, which are then usually shown at the top of the search engine result pages (SERPs). Alongside the booking of keyword ads, SEO is the second way to gain visibility in search engines [31]. Contrary to booking ads, where the advertiser pays the search engine for every click on the ad, the clicks on organic results are not associated with direct costs for the website provider. This makes SEO an attractive set of techniques to almost all content providers on the web. While search engine optimization is most often associated with optimizing for products and services, it is important to note that the same techniques are used for optimizing informational contents, i.e., users aiming to get informed about a topic will encounter optimized content (such as content produced by public-relations agencies) without knowing it.

Search engine optimization considers techniques on the website level as well as on the document level. Some measures are used to improve the overall performance of a website (e.g., improving the speed at which pages are delivered, improving the site structure). Other measures apply to the individual documents on the site (e.g., keyword density).

Search engine optimization can have positive effects. It was demonstrated that search engine optimization methods positively affect usability [55, 59] and the accessibility of websites [32]. However, it is unclear if – or to what degree – SEO also positively affects the relevance of the search results. This positive effect may stem from the fact that content is prepared to easily allow the user to find potentially relevant information objects [54]. Further, as current ranking algorithms aim to optimize user satisfaction [6], content optimized to satisfy the users will be preferred.

Overall, search engine optimization professionals paint a very positive picture of their profession, making content discoverable through search engines and bringing good content to their top results [41]. However, there is also a "dark side" to SEO, and it is unclear where the boundary between ethical and unethical work lies [66].[28] surveyed search engine optimizers, journalists, and academics, revealing the increasing importance of SEO for news outlets and, in turn, for the training of journalists. This study highlights the importance of SEO for professional content creation. Further, in an interview study with professionals at Greek media organizations, it was found that SEO increasingly influences journalists in their writing and that SEO policies are applied in newsrooms [13]. Interview studies found that journalists have reservations against SEO in the context of journalistic work [7, 13, 37]. [13] found that while SEO was considered indispensable as it ensures visibility of the content, it also has a significant influence on topic setting. This, in turn, can reduce the quality of the journalism being produced.

2.2 Selection behavior in search engine result pages

The selection behavior on the search engine results pages can be characterized as strongly oriented to the given order and representation. The most important overarching explanatory models for user behavior on the results pages are the principle of least effort [67] and satisficing [48].

As search engines present results in the form of ranked lists, the position effect plays a huge role in what users look at [23, 50] and which results they select. They predominantly select the results listed first, and an overwhelming number of clicks are made on these results. This effect has been demonstrated in numerous studies (e.g., [2, 4, 18, 19, 35, 45, 62]). However, not only the ranking of the results is important but also them being shown in the so-called visible area of the SERP, i.e., the results that are [16]. Users preferably select results from this area [20]. In current result presentations, results from multiple sources (e.g., news, videos) are shown together on a single SERP. This incorporation of results from several vertical search engines leads to a different presentation of search results. Results that take up more space on the search results page or are graphically more attractive are more likely to be perceived and selected [27].

The effect of this user behavior is shown in a large-scale study from Yahoo research [15]. Analyzing 2.6 billion search queries, this research found that in web search, about 80 percent of all clicked results are accounted for by only 10,000 websites. This underlies the massive effect of result position and the attractiveness of taking measures to make one's content be shown in these top positions.

2.3 Trust in search engines

Search engine users show a high level of trust in search engines and Google in particular. In a representative survey in the United States, three-quarters of respondents said they trusted the information they found on search engines. 28% do so for all or almost all, 45% for most information [38]. Results for the European Union are very similar: 78% of respondents in a representative survey said they trust the search results provided are the most relevant, ranging from 67% in Romania to 87% in Austria [10]. Search engine trustworthiness is comparable to traditional news media, as shown by a representative study of Internet users from 28 markets, including the United States, China, and Germany [9]. A study using a representative sample of German Internet users found that users with little knowledge of search engines are more likely to trust and use Google than other users [43].

Non-representative studies add to this picture by showing that a high ranking also increases sponsor credibility [61], that male college students seem to be more likely to trust search engines to provide objective results than women [53]. Students place search engines in terms of their results being reliable somewhere between libraries and databases (both of which were considered by most respondents to be mostly or always reliable) and internet communities, forums, blogs, and podcasts (which were considered not reliable or only somewhat reliable) [21].

In laboratory studies, it has been shown that users select results shown at the top of result lists even if they are less relevant [35, 45] or less credible [56, 58] than results shown at lower ranks.

2.4 Measuring search engine optimization

In this section, we review the literature dealing with identifying factors relevant to SEO, using them as indicators to detect SEO measures in websites or documents, and measuring the conformity of websites or documents to SEO best practices. It is important to note that the aim of the studies reported here is to measure SEO success, i.e., whether documents are optimized in ways that reflect best practices in SEO, leading to a higher ranking in commercial search engines. The focus of our research, on the other hand, is to investigate whether SEO measures have been applied to the documents, irrespective of their success.

The aim of most of the research reported here is to compare the rankings of different websites by analyzing indicators that contribute to these sites' ranking. The goal is to give recommendations on how to improve websites and documents through SEO techniques. For instance, [3] and [17] developed tools that recommend SEO measures based on analysis of a given website.

Correlational studies are based on lists of indicators that at least in some way reflect the (presumed) ranking factors of commercial search engines. The lists used in empirical work range from containing just some factors to more extensive lists [8, 17, 34, 39, 57]. Some use indicator lists even more fine-grained than the usual factors reported in the practitioner literature [34, 39]. However, these works lack a systematic collection and analysis of SEO indicators based on the literature and expert opinion.

As the studies use different sets of indicators, it is hardly possible to compare the results. Correlations between indicators and search engine rankings are found [3, 8, 11, 14, 17, 51, 57, 65], but it remains unclear which factors actually explain the rankings. Some promising factors, predominantly ones that form the basics of the professional SEO literature, are found. These include several on-page factors such as optimized meta tags (i.e., tags to provide structured metadata about a web page) and page speed optimization and several off-page factors such as the number of backlinks (i.e, links from external sites pointing to this particular page) [1, 33, 46, 64]. While factors used by SEO professionals have been applied in research, other factors indicating SEO efforts (such as the use of specialized SEO tools) have not been included so far.

Regarding the datasets used, it should be noted that it is unclear how representative the studies' queries are for the general search engine queries. Some studies also focus on specific business sectors or institutions, e.g., media [14], publishing [57], or institutions of higher education [5, 12], and, therefore, use specific sets of queries.

Further to the goal of improving rankings in search engines, [32] found positive correlations between SEO and website accessibility. [12] found correlations between quality metrics (including usability) and SEO success. This line of research indicates positive effects of SEO beyond the initial goal of achieving better ranking.

While the studies reported so far investigated correlations between result positions and SEO indicators, some research also focused on the degree to which websites or documents have been optimized. In a comparative study, [34] use a Multi-Criteria Decision Making (MCDM) algorithm to determine the degree of optimization of academic sites. [29] used machine learning algorithms to classify web pages into three predefined classes according to the degree of search engine optimization. This involved both identifying relevant features for SEO using correlation analyses and evaluating the accuracy of classification algorithms. However, a shortcoming of this study is that the accuracy is measured by comparing the algorithm's performance to expert judgments, which themselves may not be accurate. Many SEO techniques cannot be seen from a cursory inspection of a website's contents but can only be found through more in-depth technical analysis.

The literature review shows that search engine optimization is a mature line of business, applying sophisticated techniques to achieve visibility in search engines. On the other hand, search engine users choose only from a limited set of top results shown and usually do not question how these results were generated. They expose a high level of trust in search engines to provide them with the most relevant results. Prior research aiming to identify SEO in given documents focused on whether the measures taken were successful in terms of better rankings. Research gaps lie in identifying SEO in search results at a large scale, using a large number of SEO indicators, and measuring whether SEO measures have been taken (as opposed to measuring SEO success). In the remainder of this paper, we address these gaps.

3 IDENTIFYING SEO FACTORS

At first glance, one could assume that, as search engine optimizers try to reverse-engineer Google's ranking factors to make their content visible in that search engine, ranking factors and SEO factors are the same. However, we have to consider that not all SEO efforts are successful. For instance, someone trying to gain visibility in Google might use keyword stuffing, i.e., repeating a keyword on a webpage very often, suggesting to the search engine that the page is relevant to that keyword. Obviously, this approach will not work as search engines can detect such simple gaming methods. However, as we aim to detect whether content providers seek to optimize their pages, keyword stuffing may still be a factor for detecting optimized pages. Whether content providers have successfully optimized their pages is not a criterion relevant to our classification.

Our model of SEO factors and its implementation is based on an extensive review of the professional literature and interviews we conducted with SEO experts [41, 44]. In total, our model consists of 48 factors, which can be grouped along three dimensions: tools and plugins, URL lists, and indicators for SEO. We prioritized these factors for the implementation in our system. It should be noted that the current implementation considers 21 factors only. These have been considered the most fruitful by the experts and researchers, and have also been validated through machine learning methods in our initial studies. We are confident that already at this stage, we can reliably identify optimized content.

Our approach combines the automatic identification of SEO indicators from web pages and websites with manually generating lists of optimized and not optimized websites. We follow this approach as results in search engines are not equally distributed, i.e., there is only a relatively small set of websites that account for a large

fraction of the URLs shown [36] and clicked in the top results [15]. This means that by manually classifying a limited set of websites, we can already detect a relatively large fraction of optimized pages.

We created five lists of manually classified websites: SEO customers, news websites, online shops, business websites, websites with ads, and not optimized websites (for details, see Table 1). News websites are classified as being optimized as all the SEO experts we interviewed agreed that all news companies use SEO techniques to increase the visibility of their content. We manually categorized the websites by evaluating the content of 13,000 URLs. In the automatic analysis, we focused on identifying tools used by search engine optimizers on the one hand and indicators for SEO on the other hand. We differentiate between two types of tools: Tools particularly used for search engine optimization and analytics tools that are not necessarily used for SEO purposes exclusively but are usually used in the SEO context. We identified tools through analyzing the HTML code of the results found. When a tool is used, a hint can usually be found in the HTML comments or a script. The following examples show code snippets used by the Yoast SEO plugin (an SEO tool) and Google Analytics (an analytics tool):

```
<!--This site is optimized with
the Yoast SEO plugin v12.4 --
\url{https://yoast.com/wordpress} /plugins/seo/-->
<!-- Google Analytics -->
<script>(function(i,s,o,g,r,a,m)
\{i['GoogleAnalyticsObject']$=$r;$\ldots$
```

We manually extracted SEO tools and analytic tools from a set of approx. 30,000 URLs, resulting in a list of 58 SEO tool names and 54 analytics tool names, respectively. In this approach, we used lists of known SEO plugins and analytics tools[1] to check if we could find references to these plugins in the HTML comments. In addition, we searched the comments for signal words like SEO to find tools and plugins that were not included in the lists.

In terms of SEO indicators, we extract information from the page's HTML code, further information on the website level, and teste for page speed. Data extracted from the HTML code includes the use of a page description and nofollow links, among others. Information on the website level includes the use of SEO-specific information in the robots.txt file and the use of a sitemap file. Finally, we measure page speed, as one of the fundamental technical factors in search engine optimization is optimizing the pages to load quickly. We used these indicators to build a rule-based classifier to determine the probability of search engine optimization in four classes: definitely optimized, probably optimized, probably not optimized, and definitely not optimized. The rules are relatively simple since they only check the presence or absence of an indicator. A weighting was not performed. We decided to use this approach because we wanted to create a large dataset for further evaluation compared to other approaches like collecting judgments by SEO experts to build a training set [29], resulting in small datasets. Further, it is unclear how reliable experts can detect SEO measures on given webpages.

We built the dataset based on this classifier. We used it to apply machine learning methods for evaluating the rule-based classifier and to evaluate the indicators that we do not track in the rule-based approach for suitability for classifying web pages using our search engine optimization probability classes. We also investigated how well the classifier performs without features based on lists (e.g., the list with SEO tools or the list with News websites). We wanted to evaluate if we could produce useful results even for features that require constant editing and maintenance.

The dataset for our machine learning processes consisted of 281.848 documents and all of the 49 indicators. First, we performed an ANOVA-f test for feature reduction and found that we can perform our classification with 21 of the indicators. Next, we determined the importance of these features using a decision tree classifier because it gives feature importances and is a close equivalent to the rule-based approach. We created a train/test split with a test size of 0.33 (93.949 documents) using a prediction score and prediction probability score to get both the prediction from the rule-based classifier and the probability for each prediction. We then calculated the importance of the features in the classification. As a result, we used the new feature list to create a new model for a decision tree classifier to classify a different dataset. We evaluated 12 classifier algorithms (e.g., Naive Bayes Gradient Boost and Support Vector Machines) and decided to use a decision tree classifier because of the best ratio of accuracy and processing time.

The decision tree classifier performs with an accuracy of 99.7%, and a macro-precision of 99.6%, macro recall of 99.3% and a macro-f1 of 99.5% which is no surprise since we developed simple rules beforehand. The goal of this approach was to determine which features are relevant and which can be neglected. We also created a decision model without any external features, which performed with an accuracy of 79.8%. This is still good to pre-classify documents if external features from our lists are missing. In the following, we focus on the model with all features since our rule-based approach is built on it.

We built a system that automatically queries Google, collects result URLs and result positions from the SERPs, collects the result documents, analyzes the HTML code, and checks the URLs against our database of already known websites. We will not focus on the technical implementation in this paper; details can be found in [52]. Table 1 provides an overview of all indicators used for the classification. It shows the indicators from the rule-based classification with the class members, short descriptions for the rules for the classes and for the indicators, and the results from the feature reduction from machine learning. We used these results for the classification of the datasets, as detailed in the next section.

4 DATASETS

To test our system and the decision tree classifier, we used three datasets based on query sets and results collected from Google through screen scraping. Table 2 gives an overview of all datasets with a description, the number of queries, the number of scraped search results, and the source for the search queries. With various topics and a total of 256,853 result URLs without duplicates, we are confident to have built a large and diverse enough dataset for testing purposes. In the following, we will present and discuss the results for the datasets separately to show if and how the results

[1]SEO plugins: https://wordpress.org/plugins/tags/seo/ and analytics tools: https://wordpress.org/plugins/tags/analytics/

Table 1: Indicators used for classifying the documents found, ordered according to their respective rule-based class, with representation of their use in the decision tree classifier

Indicator	Description	decision tree
Definitely optimized	A result is definitely optimized if at least one of the very obvious search engine optimization criteria from our list of SEO indicators is met. Very obvious here means that the intention of SEO is clearly visible.	
SEO Tools	Tools that dedicatedly support SEO measures, e.g., Yoast SEO Plugin	x
SEO customers	Customers of search engine optimization agencies (manually collected; 1,004 items)	x
Websites with ads	List of websites showing ads (manually collected; 325 items)	x
News websites	List of news websites (manually collected; 1,203 items)	x
Microdata and schema.org	Use of microdata or schema.org on a website to define the context of the data, e.g., JSON-LD	-
Probably optimized	A result is probably optimized if it is not classified as definitely optimized, the element is not classified as not optimized, and it meets one of the indicators that we define as best practices for SEO or if the document has a visible commercial intent.	
Analytics Tools	Tools that are used for website analytics, e.g., Google Analytics	-
Online shops	List of websites (manually collected; 178 items)	x
Business websites	List of business websites (manually collected; 72 items)	-
HTTPS	Usage of Hypertext Transfer Protocol Secure	x
Pagespeed	Loading time of a website < 3s	x
SEO in robots.txt	SEO indicators in robots.txt of a website, e.g., crawl-delay	x
Nofollow links	Use of tags on the website to instruct search engines to ignore the target of the link for ranking purposes	-
Canonical links	Use of canonical tags on the website to prevent duplicate content issues	x
Online advertisements	Use of contextual and affiliate marketing on a website, e.g., Google Ads	-
Sitemap	Use of a sitemap on a website	-
Viewport	Definition of a viewport for a responsive design e.g., <meta name="viewport" content="width=device-width, initial-scale=1">	x
Open Graph Tags	Usage of open graph tags for previews of content on social media e.g., <meta property="og:title" content="website title" />	x
Probably not optimized	A result is probably not optimized if it is not definitely optimized, is not classified as not optimized, and does not have a title or description tag. These criteria are the basics of search engine optimization, so we weighted this classification result as more important if we also found criteria for classifying a result as probably optimized.	
Description	Use of a site description	x
Title	Use of a site title	x
Definitely not optimized	A result is definitely not optimized when it is on the list of definitely not optimized websites.	
Not optimized	List of websites know not to be optimized (manually collected; 1 item)	x
Features not used in the rule-based classifier		
H1 Tag	Use of H1 tags for headings on the first level	x
URL length	Length of the URL without the scheme and protocol	x
Keyword in description	Keywords of query in any description tag	x
Keyword in meta content	Keywords of query in any meta tag	x
Keyword in meta description	Keywords in the meta content tag	x
Keyword in meta open graph	Keywords in any meta open graph tags	x
Keyword in title open graph	Keywords in title open graph tag	x

Table 2: Datasets

Dataset	Description	Queries	Results	Source	Max. result output per query	Feature importance (Top-3)
Google Trends	Dataset with queries from Google Trends collected from March to June 2020 and from November and December 2020.	1,563	207,522	https://trends.google.de/trends/?geo=DE	325	News (0.53) Description (0.36) URL Length (0.15)
Radical right	Joint work with a regional media regulation authority to evaluate the use of SEO on probably radical right content.	80	12,673	Queries provided by the Medienanstalt Hamburg/Schleswig-Holstein (regional media regulation authority).	258	Description (0.56) News (0.46) Open Graph (0.18)
Coronavirus	482 queries from Germany related to the covid pandemic, collected in March 2020.	271	36,658	https://github.com/microsoft/BingCoronavirusQuerySet	277	Description (0.54) News (0.49) Open Graph (0.18)

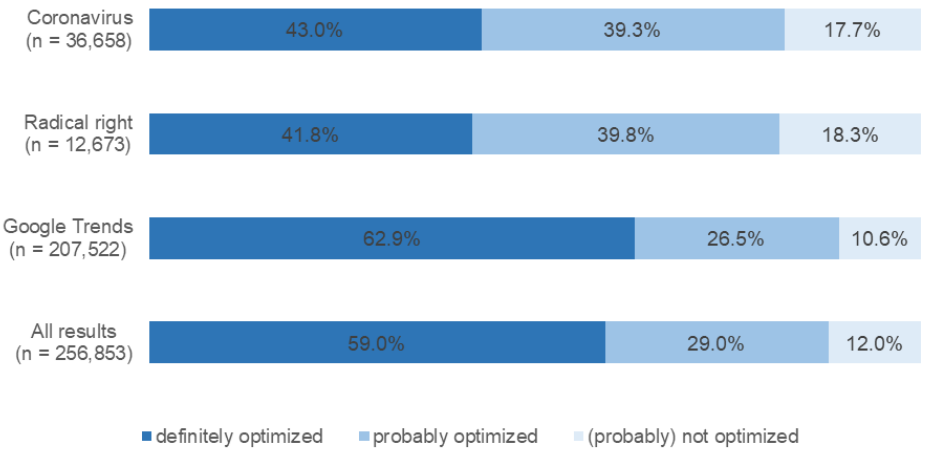

Figure 1: Results from the classification into SEO classes

depend on particular datasets and what similarities between the datasets can be found.

5 RESULTS

The results of the automatic classification are shown in Figure 1. It can be seen that the vast majority of the results are either optimized or probably optimized. Depending on the dataset, we can see that between 41.5 and 63 percent of the results found are classified as definitely being optimized. Differences between the datasets can be attributed to the higher proportion of news content found in the Trends and Corona dataset as opposed to the radical right dataset (Google Trends: 48,4 %, radical right: 27 %, Coronavirus: 31 %). Only a small fraction of the results are classified as not optimized (0.7 percent across all datasets). These are all results from Wikipedia, as this is the only website on our list of definitely not optimized sites. Our evaluation also shows that the fraction of probably not optimized results depends on the topics of the datasets. Thus, the percentage of non-optimized documents is the lowest for popular queries, at 11%, while it is around 18% for the other, thematically specific data sets.

In summary, we found that a large fraction of results found in Google is either definitely optimized or probably optimized. Over

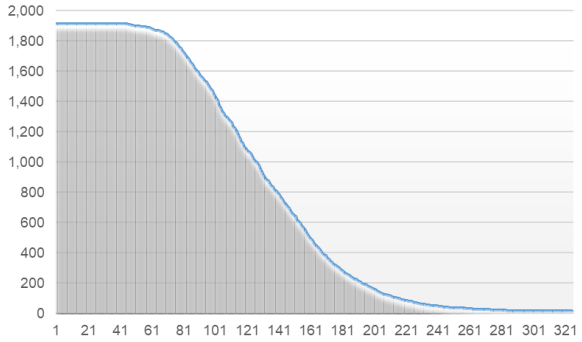

Figure 2: Results available per position (n= 1,914 queries, 256,853 documents)

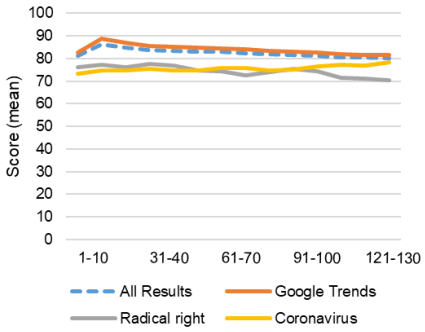

Figure 3: Score up to position 130

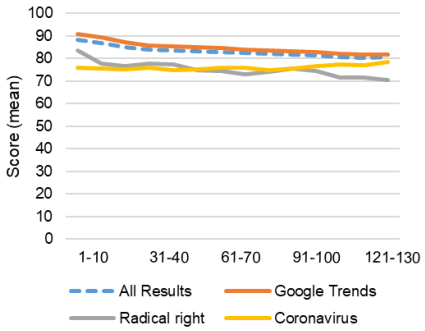

Figure 4: Score without Wikipedia up to position 130

80 percent of results found belong to these categories. This does not come as a surprise as we know that SEO is a multi-billion-dollar industry, and businesses and other actors are often dependent on the visibility their websites gain from search engine traffic. However, it should be noted that due to the limited number of results that Google provides for a query (usually not more than 300), this analysis considers these "top" results only. However, the results paint a realistic picture of what a user willing to consider all results will see, as manually querying Google will not lead to more results.

We also evaluated the probability of SEO on the top result positions across the search queries in our datasets. We decided to evaluate the score up to position 130 in Google because the number of available results decreases sharply beyond this position (see Figure 2). Thus, for many records, less than half are available from the back positions than in the top 10 results.

We translated the class affiliation to a score by defining limit values (not optimized = 0, probably not optimized = 33, probably optimized = 67 and definitely optimized = 100). Figure 3 shows the mean of the score on the positions in Google up to position 130. The mean of the score on position one is relatively low because of many Wikipedia results shown on the top position, which we always classify as not optimized. We found 1,974 results from Wikipedia in our dataset. Of these, 27 % were found on the top position and 79.3 % within the top ten positions. Figure 4 again shows the distribution, excluding Wikipedia results. Our evaluations show that the probability of SEO is slightly higher on top positions. This is especially visible on the Google Trends dataset while SEO for thematically specific topics seems to be more differentiate. SEO for content according Coronavirus is more visible on the lower rankings.

6 DISCUSSION AND CONCLUSION

This paper presented a method for identifying SEO measures in results shown by commercial search engines. Our model for identifying these measures is based on an extensive literature review, interviews with SEO professionals, and an evaluation with machine learning algorithms. The automatic classification is based on factors belonging to the three dimensions tools and plugins, URL lists, and indicators for SEO. It incorporates a total of 21 factors. An analysis based on three datasets with a total of 256,853 URLs shows that in Google, a large fraction of results available to users is optimized through SEO measures. The results indicate that the effect of SEO is stronger for popular queries.

A surprising result is that we did not find huge differences between the ratio of optimized results on different result positions. We assumed that the further one goes down the result list, the lower the ratio of optimized pages. We instead found that users will be confronted with a high number of SEO-optimized documents, even when they are willing to consider a large number of results.

This study has some limitations, the most apparent being that only data from one search engine (Google) has been analyzed. Further, while we tried to diversify queries and search results by using three different datasets, the analysis is not based on a representative sample of queries, as used by search engine users. In terms of the factors used, our analysis is limited in that we did not consider off-site factors (such as the number of backlinks). In future research, we plan to add more search engines, increase datasets in terms of size and diversity, and add more factors to our model.

In the results we were able to fetch, we found a large degree of pages being optimized. This result held for all result positions. In that sense, it would be interesting to experiment with data from search engines like millionshort.com (see [40]), which allow for removing top sources from the results and accessing results shown on positions we could not scrape from Google. We assume that in these positions, the probability of finding optimized content will be much lower.

Our research contributes to better understanding what users get to see on search engine result pages. Apart from the organic results to which the SEO measures apply, external influence can be exerted through paid search advertising (PSM; "sponsored results"), where advertisers bid for positions on the SERPs. Some research has also focused on the mixture of paid-for and organic results on search engine result pages (SERPs) and how users able vs. not able to distinguish between the two result types show a different selection behavior (e.g., [22, 24, 42]). Further, there is some research on search engine companies' self-interests and how they may influence what is shown on the result pages [25]. From the perspective of search engine providers, SEO constitutes an external influence on the ranking functions.

On the one hand, SEO benefits search engines, and search engine companies even provide help for SEO (e.g., Google's SEO Starter Guide, Webmaster Guidelines, and Search Central Help Community). On the other hand, optimized pages influence what users see in the result lists and may bias search results. This effect is further enhanced when search engines incorporate user behavior signals into ranking models through analyzing clicks and further interactions with the results (e.g., [60, 63]). This may lead to a rich-get-richer effect preferring content that is not necessarily the most relevant.

The results reported in this paper are promising, but, of course, further work is needed to refine and further evaluate the approach. In future research, it might also be interesting to bring together the different influences on the search result pages (i.e., through SEO, paid search marketing, and search engine providers' self-interests) to measure how these influence user selection behavior and knowledge acquisition through search engines.

ACKNOWLEDGMENTS

This work is funded by the German Research Foundation (DFG - Deutsche Forschungsgemeinschaft), grant number 417552432.

REFERENCES

[1] Muhammad Akram, Imran Sohail, Sikandar Hayat, M Imran Shafi, and Umer Saeed. 2010. Search Engine Optimization Techniques Practiced in Organizations: A Study of Four Organizations. *J. Comput.* 2, 6 (June 2010), 134–139.

[2] Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation bias is significant in determining user preference for search results-A user study. *J. Am. Soc. Inf. Sci. Technol.* 60, 1 (January 2009), 135–149. https://doi.org/10.1002/asi.20941

[3] Aziz Barbar and Anis Ismail. 2019. Search Engine Optimization (SEO) for Websites. In *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, ACM, New York, NY, USA, 51–55. https://doi.org/10.1145/3323933.3324072

[4] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the international conference on Web search and web data mining*, Microsoft Research, Cambridge, United Kingdom Microsoft Research, Redmond, United States, 87–94.

[5] Ashwini Dalvi and Riya Saraf. 2019. Inspecting Engineering College Websites for Effective Search Engine Optimization. In *2019 International Conference on Nascent Technologies in Engineering (ICNTE)*, IEEE, 1–5. https://doi.org/10.1109/ICNTE44896.2019.8945823

[6] Fernando Diaz. 2016. Worst Practices for Designing Production Information Access Systems. *ACM SIGIR Forum* 50, 1 (June 2016), 2–11. https://doi.org/10.1145/2964797.2964799

[7] Murray Dick. 2011. Search Engine Optimisation in Uk News Production. *Journal. Pract.* 5, 4 (August 2011), 462–477. https://doi.org/10.1080/17512786.2010.551020

[8] Ioannis C Drivas, Damianos P Sakas, Georgios A Giannakopoulos, and Daphne Kyriaki-Manessi. 2020. Big Data Analytics for Search Engine Optimization. *Big Data Cogn. Comput.* 4, 2 (April 2020), 1–22. https://doi.org/10.3390/bdcc4020005

[9] Edelman. 2020. *Edelman Trust Barometer 2020.*

[10] European Commission. 2016. *Special Eurobarometer 447 – Online Platforms.* European Commission, Brussels. https://doi.org/10.2759/937517

[11] Michael P. Evans. 2007. Analysing Google rankings through search engine optimization data. *Internet Res.* 17, 1 (2007), 21–37. https://doi.org/10.1108/10662240710730470

[12] Andreas Giannakoulopoulos, Nikos Konstantinou, Dimitris Koutsompolis, Minas Pergantis, and Iraklis Varlamis. 2019. Academic excellence,website quality, SEO performance: Is there a correlation? *Futur. Internet* 11, 11 (2019), 1–25. https://doi.org/10.3390/fi11110242

[13] Dimitrios Giomelakis, Christina Karypidou, and Andreas Veglis. 2019. SEO inside Newsrooms: Reports from the Field. *Futur. Internet* 11, 12 (December 2019), 261. https://doi.org/10.3390/fi11120261

[14] Dimitrios Giomelakis and Andreas Veglis. 2016. Investigating Search Engine Optimization Factors in Media Websites. *Digit. Journal.* 4, 3 (April 2016), 379–400. https://doi.org/10.1080/21670811.2015.1046992

[15] Sharad Goel, Andrei Broder, Evgeniy Gabrilovich, and Bo Pang. 2010. Anatomy of the long tail. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, ACM Press, New York, New York, USA, 201. https://doi.org/10.1145/1718487.1718513

[16] Nadine Höchstötter and Dirk Lewandowski. 2009. What users see – Structures in search engine results pages. *Inf. Sci. (Ny).* 179, 12 (May 2009), 1796–1812. https://doi.org/10.1016/j.ins.2009.01.028

[17] C D Hoyos, J C Duque, A F Barco, and É Vareilles. 2019. A search engine optimization recommender system. In *CEUR Workshop Proceedings*, 43–47.

[18] Thorsten Joachims, Laura Granka, Bing Pan, and Helene Hembrooke. Accurately Interpreting Clickthrough Data as Implicit Feedback.

[19] Mark T. Keane, Maeve O'Brien, and Barry Smyth. 2008. Are people biased in their use of search engines? *Commun. ACM* 51, 2 (February 2008), 49–52. https://doi.org/10.1145/1314215.1314224

[20] Diane Kelly and Leif Azzopardi. 2015. How many results per page? In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*, ACM Press, New York, New York, USA, 183–192. https://doi.org/10.1145/2766462.2767732

[21] Raphael N. Klein, Lisa Beutelspacher, Katharina Hauk, Christina Terp, Denis Anuschewski, Christoph Zensen, Violeta Trkulja, and Katrin Weller. 2009. Informationskompetenz in Zeiten des Web 2.0: Chancen und Herausforderungen im Umgang mit Social Software. *Inf. - Wiss. Prax.* 60, 3 (2009), 129–142.

[22] Dirk Lewandowski. 2017. Users' Understanding of Search Engine Advertisements. *J. Inf. Sci. Theory Pract.* 5, 4 (2017), 6–25. https://doi.org/10.1633/JISTaP.2017.5.4.1

[23] Dirk Lewandowski and Yvonne Kammerer. 2020. Factors influencing viewing behaviour on search engine results pages: a review of eye-tracking research. *Behav. Inf. Technol.* (May 2020), 1–31. https://doi.org/10.1080/0144929X.2020.1761450

[24] Dirk Lewandowski, Friederike Kerkmann, Sandra Rümmele, and Sebastian Sünkler. 2018. An empirical investigation on search engine ad disclosure. *J. Assoc. Inf. Sci. Technol.* 69, 3 (March 2018), 420–437. https://doi.org/10.1002/asi.23963

[25] Dirk Lewandowski and Sebastian Sünkler. 2013. Representative online study to evaluate the revised commitments proposed by Google on 21 October 2013 as part of EU competition investigation AT.39740-Google: Country comparison report. Hamburg.

[26] Kai Li, Mei Lin, Zhangxi Lin, and Bo Xing. 2014. Running and chasing - The competition between paid search marketing and search engine optimization. *Proc. Annu. Hawaii Int. Conf. Syst. Sci.* (2014), 3110–3119. https://doi.org/10.1109/HICSS.2014.640

[27] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of Vertical Result in Web Search Examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*, ACM Press, New York, New York, USA, 193–202. https://doi.org/10.1145/2766462.2767714

[28] Carlos Lopezosa, Lluís Codina, Javier Díaz-Noci, and José-Antonio Ontalba. 2020. SEO and the digital news media: From the workplace to the classroom. *Comunicar* 28, 63 (April 2020), 63–72. https://doi.org/10.3916/C63-2020-06

[29] Goran Matošević, Jasminka Dobša, and Dunja Mladenić. 2021. Using machine learning for web page classification in search engine optimization. *Futur. Internet* 13, 1 (2021), 1–20. https://doi.org/10.3390/fi13010009

[30] TJ McCue. 2018. SEO Industry Approaching $80 Billion But All You Want Is More Web Traffic. *forbes.com.*

[31] Mike Moran and Bill Hunt. 2015. *Search Engine Marketing, Inc.: Driving Search Traffic to Your Company's Website* (Third edit ed.). IBM Press, Upper Saddle River, NJ.

[32] Lourdes Moreno and Paloma Martinez. 2013. Overlapping factors in search engine optimization and web accessibility. *Online Inf. Rev.* 37, 4 (2013), 564–580. https://doi.org/10.1108/OIR-04-2012-0063

[33] Ushadi Niranjika and Dinesh Samarasighe. 2019. Exploring the Effectiveness of Search Engine Optimization Tactics for Dynamic Websites in Sri Lanka. In *2019 Moratuwa Engineering Research Conference (MERCon)*, IEEE, 267–272. https://doi.org/10.1109/MERCon.2019.8818903

[34] Barış Özkan, Eren Özceylan, Mehmet Kabak, and Metin Dağdeviren. 2020. *Evaluating the websites of academic departments through SEO criteria: a hesitant fuzzy*

linguistic MCDM approach. Springer Netherlands. https://doi.org/10.1007/s10462-019-09681-z

[35] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *J. Comput. Commun.* 12, 3 (April 2007), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

[36] Philip Petrescu. 2014. Google Organic Click-Through Rates in 2014 - Moz.

[37] Indra Prawira and Mariko Rizkiansyah. 2018. Search engine optimization in news production online marketing practice in Indonesia online news media. *Pertanika J. Soc. Sci. Humanit.* 26, T (2018), 263–270.

[38] Kristen Purcell, Joanna Brenner, and Lee Raine. 2012. *Search Engine Use 2012*. Washington, DC.

[39] Joni Salminen, Roope Marttila, Bernard J. Jansen, Juan Corporan, and Tommi Salenius. 2019. Using machine learning to predict ranking of webpages in the gift industry: Factors for search-engine optimization. *ACM Int. Conf. Proceeding Ser.* (2019). https://doi.org/10.1145/3361570.3361578

[40] Philipp Schaer, Philipp Mayr, Sebastian Sünkler, and Dirk Lewandowski. 2016. How Relevant is the Long Tail? In *CLEF 2016*, Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato and Nicola Ferro (eds.). Springer International Publishing, Cham, 227–233. https://doi.org/10.1007/978-3-319-44564-9_20

[41] Sebastian Schultheiß and Dirk Lewandowski. 2020. "Outside the industry, nobody knows what we do" SEO as seen by search engine optimizers and content providers. *J. Doc.* (2020). https://doi.org/10.1108/JD-07-2020-0127

[42] Sebastian Schultheiß and Dirk Lewandowski. 2020. How users' knowledge of advertisements influences their viewing and selection behavior in search engines. *J. Assoc. Inf. Sci. Technol.* (September 2020), asi.24410. https://doi.org/10.1002/asi.24410

[43] Sebastian Schultheiß and Dirk Lewandowski. 2021. Misplaced trust? The relationship between trust, ability to identify commercially influenced results, and search engine preference. *Journal of Information Science*. May 2021. https://doi.org/10.1177/01655515211014157

[44] Sebastian Schultheiß and Dirk Lewandowski. 2021. *Expert interviews with stakeholder groups in the context of commercial search engineswithin the SEO Effect project*. Retrieved from https://osf.io/5aufr/

[45] Sebastian Schultheiß, Sebastian Sünkler, and Dirk Lewandowski. 2018. We still trust in google, but less than 10 years ago: An eye-tracking study. *Inf. Res.* 23, 3 (2018).

[46] Jenna Pack Sheffield. 2020. Search Engine Optimization and Business Communication Instruction: Interviews With Experts. *Bus. Prof. Commun. Q.* (January 2020), 232949061989033. https://doi.org/10.1177/2329490619890335

[47] Similarweb. 2021. SimilarWeb | Website Traffic Intelligence.

[48] Herbert Alexander Simon. 1955. A Behavioral Model of Rational Choice. *Q. J. Econ.* 69, 1 (1955), 99–118. https://doi.org/10.2307/1884852

[49] StatCounter. 2020. Search Engine Market Share Europe | StatCounter Global Stats.

[50] Artur Strzelecki. 2020. Eye-Tracking Studies of Web Search Engines: A Systematic Literature Review. *Information* 11, 6 (June 2020). https://doi.org/10.3390/info11060300

[51] Ao-Jan Su, Y Charlie Hu, Aleksandar Kuzmanovic, and Cheng-Kok Koh. 2014. How to Improve Your Search Engine Ranking: Myths and Reality. *Acm Trans. Web* 8, 2 (2014), 8. https://doi.org/10.1145/2579990

[52] Sebastian Sünkler and Nurce Yagci. 2021. Development and software implementation of a preliminary model to identify the probability of search engine optimization on webpages. Hamburg.

[53] Arthur Taylor and Heather A. Dalal. 2017. Gender and Information Literacy: Evaluation of Gender Differences in a Student Survey of Information Sources. *Coll. Res. Libr.* 78, 1 (2017), 90–113. https://doi.org/10.5860/crl.78.1.90

[54] Shari Thurow. 2015. To Optimize Search, Optimize the Searcher. *Online Search.* 39, 4 (2015), 44–48.

[55] Shari Thurow and Nick Musica. 2009. *When Search Meets Web Usability*. New Riders, Berkeley.

[56] Andreas Tremel. 2010. Suchen, finden - glauben? Die Rolle der Glaubwuŕdigkeit von Suchergebnissen bei der Nutzung von Suchmaschinen. Ludwig-Maximilians-Universität (LMU) München.

[57] Lance Umenhofer. 2019. Gaining Ground: Search Engine Optimization and Its Implementation on an Indie Book Press. *Publ. Res. Q.* 35, 2 (June 2019), 258–273. https://doi.org/10.1007/s12109-019-09651-x

[58] Julian Unkel and Alexander Haas. 2017. The effects of credibility cues on the selection of search engine results. *J. Assoc. Inf. Sci. Technol.* 68, 8 (August 2017), 1850–1862. https://doi.org/10.1002/asi.23820

[59] Eugene B. Visser and Melius Weideman. 2011. An empirical study on website usability elements and how they affect search engine optimisation. *SA J. Inf. Manag.* 13, 1 (March 2011), 1–9. https://doi.org/10.4102/sajim.v13i1.428

[60] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. *SIGIR 2016 - Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* (2016), 115–124. https://doi.org/10.1145/2911451.2911537

[61] Axel Westerwick. 2013. Effects of Sponsorship, Web Site Design, and Google Ranking on the Credibility of Online Information. *J. Comput. Commun.* 18, 2 (January 2013), 80–97. https://doi.org/10.1111/jcc4.12006

[62] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias. In *Proceedings of the 19th international conference on World wide web - WWW '10*, ACM Press, New York, New York, USA, 1011. https://doi.org/10.1145/1772690.1772793

[63] Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. 2017. Situational context for ranking in personal search. *26th Int. World Wide Web Conf. WWW 2017* (2017), 1531–1540. https://doi.org/10.1145/3038912.3052648

[64] Lihong Zhang, Jianwei Zhang, and Yanbin Ju. 2011. The research on Search Engine Optimization based on Six Sigma Management. In *2011 International Conference on E-Business and E-Government (ICEE)*, IEEE, 1–4. https://doi.org/10.1109/ICEBEG.2011.5881880

[65] Christos Ziakis, Maro Vlachopoulou, Theodosios Kyrkoudis, and Makrina Karagkiozidou. 2019. Important factors for improving Google search rank. *Futur. Internet* 11, 2 (2019). https://doi.org/10.3390/fi11020032

[66] Malte Ziewitz. 2019. Rethinking gaming: The ethical work of optimization in web search engines. *Soc. Stud. Sci.* 49, 5 (2019), 707–731. https://doi.org/10.1177/0306312719865607

[67] George Kingsley Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. https://doi.org/10.2307/2226729

APPENDICES

A RESEARCH DATA

Research data and software code are available at https://doi.org/10.17605/OSF.IO/JYV9R.