Comparison of Results Shown by Different Search Engines for Climate-Related Topics

Bachelor thesis to obtain the degree

Bachelor of Science

submitted to the Faculty of Engineering
Department of Computer Science and Applied Cognitive Science
Prof. Dr. Dirk Lewandowski

in the bachelor's program
Computer Engineering (Software)-ISE

Author	Tuhina Kumar
Student ID Number	3085926
First examiner	Prof. Dr. Dirk Lewandowski
Second examiner	Prof. Dr. Nobert Fuhr
Editing period	07.03.2023 – 07.06.2023

Abstract

This thesis presents a comparative analysis of search engine results obtained from Google, a prominent commercial search engine, and Ecosia, an environmentally-focused search engine that reinvests its profits into tree planting initiatives. The aim of this thesis was to investigate whether Ecosia, with its environmental mission, demonstrates different search results compared to Google. The analysis focused on a dataset comprising climate-related search queries and explored the distribution of domain categories, the positioning of search results, and the level of overlap between the two search engines. The analysis revealed that there was an overall overlap of 25% between the two search engines. Interestingly, both Google and Ecosia prominently displayed governmental and intergovernmental organizations as the most frequently appearing domains. Despite Ecosia displaying a more diverse range of categories, the anticipated outcome of showing a greater number of climate-related or non-profit domains did not materialise. The limitations of this analysis were the narrow scope of the data, which was collected from a single location and focused on queries from only three broad climate themes: Global Warming, Greenhouse Effect and Ozone Layer Depletion. This can be addressed in future research by incorporating a wider variety of query themes and diversifying the locations where data is collected. In addition, it would be beneficial to evaluate the relevance of search results from Google and Ecosia, even when the overlap between the search engines is minimal.

Table of Contents

List of Figures

List of Tables

Introduction

In today's world, information is considered one of the most important assets, making it a necessity that relevant information for any decision-making process be efficiently made available. Search engines are crucial platforms, and they enable a value exchange between three types of users: information-seeking consumers, advertisers attempting to reach those consumers, and content providers (Erdmann et al., 2022). As of 2019, four billion people are internet users, and one million new users sign up every day, claim Jungblut and Stolz (as cited in Sharma & Dash, 2022). This figure represents more than half of the global population. The increasing number of internet users has revolutionised the way people communicate, access information, and conduct business. It has also created new opportunities for innovation and growth in various industries.

As technological advancements get more sophisticated and technology becomes an indispensable part of our everyday lives, its adverse impacts are scrutinised more than ever. One of the most pressing ethical issues we face today is climate change, which is caused by human activities and has far-reaching consequences for the planet. According to the UN's Environment Programme, the IT sector currently accounts for about 2–3% of global emissions (2021). Internet use is responsible for a significant portion of these carbon emissions (Sharma & Dash, 2022).

It is estimated that the emission of greenhouse gases for the average search query on Google fluctuates between 0.2 g and 10.0 g of CO2 (Glass, 2009; Hölzle, 2009; Leake and Woods, 2009; as cited in Ruch et al., 2011). Jens Gröger, a senior researcher at the Öko- Institute, estimates that a single search query yields 1.45 grams of carbon dioxide emissions (Jungblut & Stolz, 2019). Considering that a person runs 50 search

queries per day, they would produce about 26 kg of CO2 annually (Jungblut & Stolz, 2019). These emissions may seem insignificant, but when multiplied by the billions of daily internet users worldwide, they have a significant impact on the environment. In 2016, Google was estimated to have a carbon footprint of 2.9 million tons (Jungblut & Stolz, 2019).

Relevance of the Topic

The lack of a long-term strategic analysis of keyword selection translates to organic traffic relying on findings from Search Engine Marketing (SEM) literature (Erdmann et al., 2022). Therefore, the aim of this thesis was to present and compare the results from two search engines for searches in the area of climate-related themes.

It is quite clear that the technology sector has a significant carbon footprint. The high energy consumption has led to concerns about the environmental impact of data centres, as well as the cost of electricity for companies that operate them. To address these concerns, some companies are exploring more sustainable and energy-efficient practices. But despite an increase in the use of renewable energy, only a small portion of the electricity used by data centres comes from renewable sources (Sharma & Dash, 2022). Therefore, as a direct response to the environmental and social impact of deforestation, Christian Kroll founded Ecosia in 2009 (Ecosia GmbH, 2023). The German-based company has developed a carbon neutral search engine and follows a social business model. 80% of the advertising revenue is donated to tree planting projects and over 160 million trees worldwide have been planted as of 2023 (Ecosia GmbH, 2023), thus making it a prime example of how Green IS can be used for social and environmental good. This not only helps to mitigate the negative impact of technology on the environment but also promotes sustainable development.

Given the environmental focus of the thesis, it was particularly intriguing to examine the potential differences in search results and rankings across different search engines, as understanding how different search engines handle climate-related themes is crucial for evaluating their relevance in providing accurate and comprehensive information.

Structure of the Work

When speaking of search engines, Google is the first thing that comes to people's minds, and that is backed by the fact that over 90% of all the queries in Germany are made to Google (Lewandowski, 2021). Owing to its high user base and popularity, Google was the one of the two search engines for which the search results were collected and compared.

Other search engines being used as alternatives include Bing, Yahoo!, DuckDuckGo, Baidu, Yandex and Ecosia. Since climate related search results were investigated, Ecosia, a CO2-negative search engine, was picked for statistical comparison against Google. It was also noted that Ecosia is a part of the Microsoft Search Network and therefore works on the same algorithm as the second most popular search engine-Bing (Ecosia, 2023). Thus it may be concluded that the search results of Bing and Ecosia will be similar., due to which Bing was logically eliminated. Also since Ecosia boasts of privacy features, like encrypted searches, and search anonymization, ensuring user data is kept safe (Ecosia GmbH, 2023), its efficacy vis a vis Google will evince interest from users conscious of their privacy.

Climate change covers a broad spectrum of topics and subtopics, hence, for the purposes of this thesis, the following themes were selected: Global Warming, Greenhouse Effect and Ozone Layer Depletion.

Literature Review and State of Art

Information Retrieval and Search Engine Algorithms

The ranking of search results in an information retrieval system stems from the assumption that information-seeking users should only receive information relevant to their search query (Behnert & Lewandowski, 2015). Additionally, studies have shown users' tendencies to browse only through the first few results, making ranking results a crucial step towards the success of a search engine. (Enge et al., 2015; Segev, 2010; Silverstein et al., 1999). This explains why webpage authors compete for higher rankings within engines (Bar-Ilan et al., 2006) and why search engine optimization is a flourishing domain.

Search engine results can vary depending on multiple factors, including location, user personalization, and algorithm variations. Users have a significant impact on the design of search engines because they use data generated in the ranking of search results and the layout of the user interface, implying that search engine design is influenced not only by the developers but also by the users of the engines (Lewandowski, 2021). When retrieving information from the internet, the speed and relevance of the search results play a significant role in determining which search engine to employ.

However, relevance is subjective and user dependent (Maillé et al., 2022). A possibility to facilitate the relevance of results is personalising them with the help of cookies and user location (Maillé et al., 2022) . This approach, however, has the disadvantage of producing biased results (Maillé et al., 2022), which arise from the tendency of a search engine to favour particular results through the assumptions built into its algorithms (Lewandowski, 2017). A balanced and neutral representation of

results becomes especially important in cases where the query is potentially controversial and can have opposing views (Gezici et al., 2021).

User Behaviour and Search Engine Usage

As demonstrated by the billions of clicks that are generated on search engines every day, online searches are unquestionably the most heavily used resource for information retrieval (Wang & Pleimling, 2017; Wellings & Casselden, 2019). User behaviour and search engine usage have evolved over the years, with users becoming more sophisticated in their search queries and expectations. Search engines have responded by constantly improving their algorithms to provide more accurate and relevant results. While early search engines relied on crude criteria such as the number of page visits or back links, modern search and ranking algorithms are far more sophisticated and hidden from view (Hinman, 2018). A website showing up on top of the search result can no longer be credited merely to its popularity or the number of websites that link to it (Hinman, 2018). It has multiple parameters, such as location, web history (Wellings & Casselden, 2019), and more complex algorithms at play (Yu, 2016, as cited in Wellings & Casselden, 2019).

Research conducted by Wellings and Casselden in 2019 found that a fair share of people do not have a basic understanding of how search engines operate and provide results, yet it is an indispensable tool in their lives and is used heavily on an everyday basis. They used Google as an example, stating that it is not impartial and that getting the same search results on two different computers is unlikely, whereas more than half of the studied group, which consisted of scientists and engineers, believed otherwise (Wellings & Casselden, 2019) . Despite not knowing the inner workings behind it, a staggering majority of over 70% believed that search engines returned the most

relevant results based on their query, and over 60% said they trusted the information they found online. (Wellings & Casselden, 2019) . The trust that users have in the accuracy and relevance of search engine results makes it clear that technology has a clear influence in the lives of people. As a consequence, it highlights the ethical responsibility of search engines, making it important for search engines to give due importance to the accuracy and reliability in search results, as well as ensuring the avoidance of bias (Lewandowski, 2017). It is essential to understand user behaviour in terms of search engine usage to analyse search engine data, as it helps get a better understanding of user expectations and preferences as well as the search trends at any given time.

A study conducted by Sisco et al. in 2021 using data on internet searches run by users in 46 different countries found that, compared to previous years, the amount of information sought about climate change significantly increased in 2019. This rise in interest can be attributed to a number of factors, including governments and organisations around the world starting to take more actions to combat climate change and raising awareness through political events like the yearly United Nations Climate Change Conference (Sisco et al., 2021). The increasing awareness of the devastating impact that climate change is having on our planet can additionally be attributed to climate-related events and marches (Sisco et al., 2021), as well as the increased coverage of these events in the media. According to research, witnessing acts of social activism can evoke powerful emotions in viewers (Landmann & Rohmann, 2020, as cited in Sisco et al., 2021), further leading to more people trying to get informed on the topic, resulting in increased searches (Sisco et al., 2021). Sisco et al. propose that climate activism has a direct and significant influence on information seeking (2021).

Although global warming and climate change are long term trends, often unrelated to the local temperature, local weather fluctuations are a common source of climate-related searches (Choi et al., 2020; Sisco et al., 2021). Choi et al. explain that this is due to the fact that the local weather is the element that has a direct impact on people's lives and experiences (2020). Their study further validates this with the finding that in months that are warmer than usual in a particular city, the number of Google searches for the term "global warming" increases (Choi et al., 2020).

Reyes-Menendez et al. indicate a strongly positive correlation between the success of a search engine and the chance given to users to make a social or environmental impact by a search engine (2018). For this reason, some people might prefer using Ecosia, given that it uses its profits to plant trees and supports reforestation projects (Ecosia GmbH, 2023). A unique consumer group, LOHAS (Life of Health and Sustainability), has emerged in recent years, which comprises individuals who prioritize social and environmental responsibility in their purchasing decisions (Choi & Feinberg, 2021). According to the findings of a survey that was carried out by Allianz on people aged 14 and older in Germany, 44% of Germans consider themselves to be LOHAS consumers (Allianz Deutschland AG, 2009). As a result, they may be more inclined to use search engines with a social business model, like Ecosia, that align with their values and beliefs instead of a conventional one. (Schmidt, 2011) .

Alternative Search Engines

According to the data collected by StatCounter (2023) between March 2022 and March 2023, the distribution of the market shares of search engines in Germany, as well as across the world has been listed in table 1.

Table 1

Comparison of market shares (StatCounter, 2023) held by different search engines in Germany and across the world.

Market Share	Search Engine					
	Google	Bing	Yahoo!	DuckDuck Go	Ecosia	Yandex
Germany	90.05%	5.69%	1.16%	0.93%	0.82%	0.74%
Worldwide	93.18%	2.87%	1.12%	0.52%	0.09%	1.02%

It should come as no surprise that Google, regardless of the region chosen, is the undisputed leader in the search engine market. According to Yagci et al., one search engine may be heavily preferred over the others due to factors such as usability, specialised features, and easier integration into their technical environment (2022).

Google's dominance can be attributed to its sophisticated algorithms, which deliver highly relevant search results quickly and accurately. Additionally, Google's user-friendly interface and extensive range of features (Parsania et al., 2016) such as Google Maps, Gmail, and Google Drive, make it a preferred choice for many users. It has further strengthened its position as a user favourite by integrating all of their products (Parsania et al., 2016) to create an easy-to-use ecosystem for its users.

Google might provide accurate, customised results, but it comes at the cost of the privacy of the user (Krupp & Gersey, 2022). The retention of user data provides search engines with multiple long-term benefits, which in turn aids in their long-term ability to maintain their market share (Chiou & Tucker, 2017). While some users may feel that the convenience and advantages of customization outweigh the potential risks, this is a major concern for users who value their privacy and want to keep their personal information secure.

Increasing awareness about the importance of data security has led some users to seek out more privacy-focused search engines, such as Ecosia. Ecosia focuses on being environment friendly by donating a portion of its ad revenue towards planting trees globally. The search engine also boasts privacy features, like encrypted searches, and search anonymization, ensuring user data is kept safe (Ecosia GmbH, 2023).

Although several new search engines are entering the market, their actual number is overstated, due to the fact that a number of search engines only serve as search channels by displaying results from a partner rather than producing them from their own index (Yagci et al., 2022).

Methodologies for Comparing Search Engine Results

According to Bar-Ilan et al., the objective of the algorithmic ranking functions of search engines is to rank the "most relevant" results first (2006). However, with relevance being a rather subjective concept with no clear definition, it is a problematic notion. It largely depends on the users' information seeking objectives, with some researchers suggesting that the only judge of the relevance of the results is the user with the information problem (Gordon & Pathak, 1999, as cited in Bar-Ilan et al., 2006), making human judgement a viable method of relevance assessment. This is achieved via query collection followed by jury evaluation.

In order to avoid learning and branding effects, the results are collected, anonymized, and mixed at random before being evaluated by jurors for relevance, according to Lewandowski (2015). The juror-based study with anonymised results is followed by the reassignment of the results to their respective search engines, creating a sample ready to be analysed (Lewandowski, 2015).

There are multiple ways to prepare an efficient test for jurors. The majority of research utilizes three- to six-point relevance scales or binary relevance decisions, which is a straightforward approach, making it versatile for calculations of measures like the arithmetic mean (Lewandowski, 2015). Some studies employ this methodology with the added option of designating a result as one that points towards a relevant result (Griesbaum, 2004; Griesbaum et al., 2002, as cited in Lewandowski, 2015). However, according to Lewandowski, this approach is problematic as it can potentially require a user to go through multiple links before finding a relevant result (2015).

Other studies, like the one conducted by Beg in 2005, made use of user behaviour like the order in which the documents are visited and the time spent viewing them (as cited in Bar-Ilan et al., 2006). The majority of these studies, however, employ a single juror for each query (Lewandowski, 2015), which is a limitation that could lead to inaccurate results given that a result deemed relevant by one user may be deemed irrelevant by another. (Huffman & Hochster, 2007, as cited in Lewandowski, 2015).

There are other methods for evaluating search results that do not involve user participation. Zhao carried out one such study in 2004 in which they looked at changes in the rankings of the 24 websites that were among the top 20 pages throughout the data collection period by submitting the search term "cataloging department" to Google once a week for a period of 10 weeks (as cited in Bar-Ilan et al., 2006).

Alternative approaches for evaluating search results involve using search engine data to assess the size of overlap and utilising a diverse range of measures. These measures include Spearman correlation, which quantifies the relationship between overlapping

elements, a normalised Fargin measure (Bar-Ilan et al., 2006), the Jaccard index, and the Gini coefficient (Yagci et al., 2022). These evaluation techniques provide insights into the similarity, diversity, and ranking of search results, enabling a broad assessment and comparison of search engines.

Material and Methods

Data Procurement

Data is an essential part of any study. There are various approaches to extracting data from search engines, some of which are more labour-intensive than others. Certain search engines support automatic querying via application programming interfaces (APIs). According to Lewandowski & Sünkler, the main drawback to this approach is that the results are different from those that are displayed on the user interface (2013).

Another useful approach is web scraping, which resolves the above mentioned problem. It is a different form of data mining and can be considered a practical technique for extracting unstructured data from websites and converting it into a structured format (Sirisuriya, 2015). This is done in standard structures, like spreadsheets, databases, or a comma-separated values (CSV) file that can be stored and analysed in a database (Sirisuriya, 2015). Screen scraping, as Sünkler et al. explain, extracts result descriptions, result URLs, result types, etc. from the search engine results pages (SERPs), which are the same results also visible to a normal search engine user (Lewandowski & Sünkler, 2013) .

For this thesis, a web scraping software, called the Relevance Assessment Tool (RAT) was used. It is a versatile, web-based software that can be used to effectively retrieve data from commercial search engines (Lewandowski & Sünkler, 2019). The HTML source code of the search engine result page (SERP) can be used to obtain information such as the hit description, the URL, the hit position, and so on (Lewandowski & Sünkler, 2019). It has the added benefit of being able to design studies with questions and scales, and also use the individual components for a multitude of studies

(Lewandowski & Sünkler, 2019). It can also be used for juror evaluations, and the rapid crowdsourcing speeds up the collection of many evaluations, greatly improving the data collection process (Lewandowski & Sünkler, 2019).

Various studies have shown that the majority of the clicks are made on the first ten search results (Enge et al., 2015). It was therefore decided that only the first ten search results displayed by Google and Ecosia would be analysed. Statistical comparison between the results from the two search engines was used to determine if Ecosia, a search engine which is marketed as an initiative towards the preservation of the environment, displayed dissimilar results.

A query set was decided upon, and it included three broad themes, namely global warming, greenhouse effect, and ozone layer depletion. This list was inputted into the Relevance Assessment Tool, which aided in the generation of search results for both search engines for further analysis. The output, which was the raw dataset, was received in both CSV and XLSX formats.

Data Cleaning and Processing

The initial inspection of the raw data led to the identification of errors and inconsistencies. To address the issues, the procured raw data was methodically cleaned as a subsequent step to ensure the accuracy and reliability of the analysis.

In the domain column, some entries were invalid and were leading to inaccurate results. The IP address for all invalid entries was observed to be 0.0.0.0. Microsoft Excel was employed to apply a filtering process to eliminate rows containing erroneous IP addresses, resulting in the removal of 195 out of the total of 48457.

Additionally, Bing appeared as the eighth most common domain on Ecosia, with a frequency of 756 occurrences. As Ecosia sources its results from Bing (Ecosia, 2023), this was seen as an interesting observation. On closer inspection, it was discovered that the URLs for all of these results were broken, possibly as a result of an error that occurred when the data retrieval tool attempted to retrieve data from Ecosia. These 756 rows were therefore deemed redundant and removed from the dataset too.

Next, the total number of rows for the dataset differed when viewed on Excel versus when viewed with the aid of Pandas. The Excel file was examined for hidden formatting and filtering, but this was ineffective. The presence of blank rows in the data set was another potential cause. When reading the file in Pandas, the dropna() function was utilised to eliminate all rows without values, which resolved the issue, leaving behind 47506 final rows.

Furthermore, having columns in the data set that are not being used leads to storage waste and a potential increase in the runtime of data processing. To address this issue, a preliminary analysis of the raw data was done, the necessary columns relevant to the analysis were selected as parameters, and the remaining columns were removed.

Computational Setup

This section provides an overview of the computational setup and the steps that were performed in the preparatory stages of the analysis. The analysis was conducted using the programming language Python, and Jupyter Notebook was selected as the working environment. Python version 3.11.3 was used as it was the latest version and was compatible with the required libraries. Jupyter Notebook version 6.0.3 was used, as it

provided an interactive platform for efficiently executing code and viewing the results.

The Pandas library (version 1.5.3) was extensively utilised for data manipulation and processing tasks. The NumPy library (version 1.24.3) provided support for dealing with multiple numerical operations and statistical analysis efficiently. The Matplotlib library (version 3.7.1) was used to visualise the data in the form of various types of graphs and charts. The openpyxls library (version 3.1.2) helped in the seamless interaction with Excel files for data import, export, and transformation.

To ensure reproducibility and maintain a consistent environment, a virtual environment was created using Conda, a package and environment management tool within Anaconda. The virtual environment was set up with the Python version and library dependencies required for the analysis.

The code was organised into separate cells, allowing for improved accessibility and readability by facilitating the division of code blocks into distinct and manageable sections. The implemented algorithms included data loading, cleaning, transformation, and various statistical analyses. Comments were incorporated within the code to provide clarity and understanding.

By following these steps, a reliable experimental setup was established, allowing the analysis to be conducted in a controlled and reproducible manner.

Comparison Methods

The study was carried out in a series of steps. The methodology used in the study conducted by Yagci et al. in 2022 was used as reference for this analysis. As a first

step, the data was split into separate dataframes, based on the search engine, for making it easily accessible for further analysis.

Due to the theme of the queries having a narrow focus, manual classification of the domains was done. This was accomplished by first creating a broad set of categories based on the top 200 unique domains found across the results, and then manually classifying these domains under the decided 9 categories, as listed in table 2.

Table 2

The categories for domains with their description

Category	Description	Example
Governmental and Intergovernmental Organisations	Official and verifiable websites affiliated with one or more governments	EPA, NASA, UN
Encyclopedia	Websites that provide exhaustive information on a wide range of subject categories, including dictionaries.	Wikipedia, World Atlas, Merriam Webster
Journals and Research Institutions	Websites related to scientific journals, academic publications, research institutions, and educational institutes	Royal Society, ResearchGate, Stanford
Non-profit Organisations and Climate Activists	Non-profit organizations and climate activist websites advocating for sustainability, environmental preservation, and climate change. An organization's profitability was an exclusion criterion.	WWF, GreenPeace, UCS USA
Media and News Outlets	News agencies and media outlets, which were verifiable and reliable information sources, were deemed trustworthy.	BBC, DW, National Geographic

Category	Description	Example
Social Media and Blogs	Social media websites, and personal and professional blogs.	Reddit, Treehugger
Educational Platforms and Learning Resources	Domains containing educational resources, online courses, and exam preparation material	GeeksforGeeks, Byjus
Private Companies	Domains of for-profit organizations that offer services with the intention of generating a profit.	NationalGrid, Iberdrola, Amazon
Miscellaneous	Domains that fit into none of the above categories	World Economic Forum (WEF)

The top 10 domains from both search engines were retrieved based on the frequency of their occurrence. Instead of using the entire URL, the "Main" column of the dataset was used, which provided the isolated main domain. The count of their occurrences, in addition to their mean position and the standard deviation of their positions, was calculated.

In statistics, the arithmetic mean, as Rosenkrantz describes, is the ‘center of mass of a distribution’, while the standard deviation is the measure used to display the degree of dispersion of the actual values from the mean position of the values (2008). In terms of search engine results, the application of the standard deviation as a measure of dispersion for a ranking list has proven to be an effective technique (Pérez-Iglesias & Araujo, 2010). The combination of the mean position and the frequency of appearance of a domain served as a good indication of how relevant or important a

domain is considered by the search engine. For this analysis, both measures were calculated using the ‘mean’ and ‘std’ functions of the numpy module, respectively.

With the aid of the previously performed category classification, the determination of the distribution of the types of domains, based on position was done. The top few results are the ones receiving the maximum clicks (Enge et al., 2015), and users typically trust the order of search engine results for relevance (Guan & Cutrell, 2007), suggesting that a user who reaches a lower rank would already have visited the higher-ranked links. In order to analyse the category distribution, the data for positions were, therefore, evaluated cumulatively, along with all positions ranked higher than the positions in question. This was represented by a grouped bar chart depicting the percentage share of each category based on the cumulative position data.

The Jaccard similarity index (González et al., 2008, as cited in Yagci et al., 2022) for the two search engines was also calculated. The use of this index to measure the similarity between two data sets as a single number between 0 and 1 was demonstrated in a 2019 study conducted by Puchmann (Yagci et al., 2022). 0 represented total dissimilarity, whereas 1 indicated that the sets were identical. It was computed by dividing the intersection of the two sets by their union.

In the next step, the results generated by each query were compared. The number of identical results, out of ten, was determined by comparing the domains displayed by both search engines in response to the search query. In order to better visualise the results, the cumulative frequency distribution of a specific number of common domains was represented as a histogram, which is a series of vertical bars . The arithmetic mean was calculated to determine the average number of shared domains, while the median was determined to gain a better understanding of the distribution.

Results

Data Description

As displayed in table 3, the data extracted from the RAT contained a total of 47506 rows, of which 23908 were for the first search engine, Google, and 23598 for the second search engine, Ecosia. A total of 2459 queries were looked up in both search engines, and the top 10 results from both were extracted. For the results of the queries in both search engines, a total of 2442 unique websites were found. Google returned 1934 unique domains in response to the queries, while Ecosia returned 1127. The two search engines returned a total of 619 unique domains that were common to both.

Table 3

Description of the Data Set with the Count of Different Elements

	Count	
	Google	Ecosia
SE-specific Rows in Dataset	23908	23598
SE-specific Unique Domains	1934	1127
Exclusive Domains	1315	508

	Count
Total Rows in Dataset	47506
Total Queries	2459
Total Unique Domains	2442
Common Domains	619

Popular Domains

Table 4 shows the results from the comparison of the popular domains of Google and Ecosia. When assessing the top ten most popular domains across the two search engines' combined search results, climate.nasa.gov was the only domain to appear in the top three of both search engines. However, it had a mean position of 3.33 and a standard deviation of 2.69 in Google, compared to 4.59 and 2.70 for Ecosia. Additionally, it was observed that the top domain in Ecosia had almost twice as many occurrences at 1901, compared to the top domain in Google, which had 1056. The

mean position of the top domain was seen differing again in both search engines, with Google displaying EPA at a mean position of 3.89 and Ecosia showing Britannica at a mean of 4.96. The standard deviations for both were quite similar, with 2.90 and 2.69 for Google and Ecosia, respectively. Other websites that appeared in the top ten popular results for both search engines included epa.gov, nationalgeographic.org, britannica.com, and un.org, but in varying mean positions.

In addition to the different mean positions of the common websites that were displayed, there were notable differences between the two search engines when comparing the categories of websites displayed as well. The most frequently displayed domain for Google fell under the category “Governmental and Intergovernmental Organisations” (see Table 4). For Ecosia, this spot was taken by the “Encyclopedia” category. However, on a closer examination, it was observed that National Geographic appeared twice, the occurrences of which, if combined, added up to 2551, which was well above that of the domain listed in the first position, Britannica, which had 1901 (see Table 4).

The discrepancy was due to the domains having different extensions. When accounted for, this made the top domain for Ecosia fall under the category of “Media and News Outlets” (see Table 4). Both instances of National Geographic also had a slightly higher mean position than Britannica.

Both search engines prominently displayed results from the categories "Governmental and Intergovernmental Organizations” and "Encyclopedia". The Google results also featured NRDC, which was a heavily searched domain with 805 occurrences and fell under the "Non-Profit Organizations and Climate Activists" category. This category was not represented in Ecosia's top ten domains.

Table 4

Top 10 domains (URL) for Google and Ecosia on the basis of their count with their respective count of occurrence (#), Mean Position (M), Standard Deviation of Position (SD) and the Category (C) that the domain falls under.

	Google					Ecosia				
No.	URL	#	M	SD	C[a]	URL	#	M	SD	C[a]
1	www.epa.gov	1056	3.89	2.90	Gov	www.britannica.com	1901	4.96	2.69	Ency
2	en.wikipedia.org	1038	4.39	2.57	Ency	climate.nasa.gov	1830	4.59	2.70	Gov
3	climate.nasa.gov	810	3.33	2.69	Gov	www.nationalgeographic.org	1419	4.06	2.68	News
4	www.nrdc.org	805	5.24	2.46	NP	www.nationalgeographic.com	1132	4.55	2.32	News
5	www.britannica.com	659	6.22	2.29	Ency	www.epa.gov	1119	5.98	2.66	Gov
6	www.un.org	580	4.83	2.91	Gov	en.wikipedia.org	958	4.80	2.98	Ency
7	www.bgs.ac.uk	542	5.12	2.47	Res	www.un.org	846	4.78	2.38	Gov
8	public.wmo.int	517	4.97	2.84	Gov	climatekids.nasa.gov	557	5.06	2.30	Gov
9	byjus.com	510	5.88	2.75	Edu	www.bbc.co.uk	542	6.84	2.62	News
10	www.nationalgeographic.org	506	5.80	2.64	News	www.bbc.com	487	5.37	2.89	News

Note. [a]The categories mentioned in the table are Governmental and Intergovernmental Organizations (Gov), Encyclopedia (Ency), Journals and Research Institutions (Res), Non-profit Organisations and Climate Activists (NP), Media and News Outlets (News), Social Media and Blogs (SoM), Educational Platforms and Learning Resources (Edu), Private Companies (Pvt) and Miscellaneous (Misc)

The standard deviation for the majority of domains tended to range between 2 and 3 on average. Several domains were discovered to have an exceptionally low standard deviation in instances where their occurrence was also negligibly low. The highest standard deviation witnessed for both Google and Ecosia was 6.36, however Google had 11 domains that had a standard deviation higher than five with a total occurance of 22, while Ecosia had only 3 with a total occurance of 6.

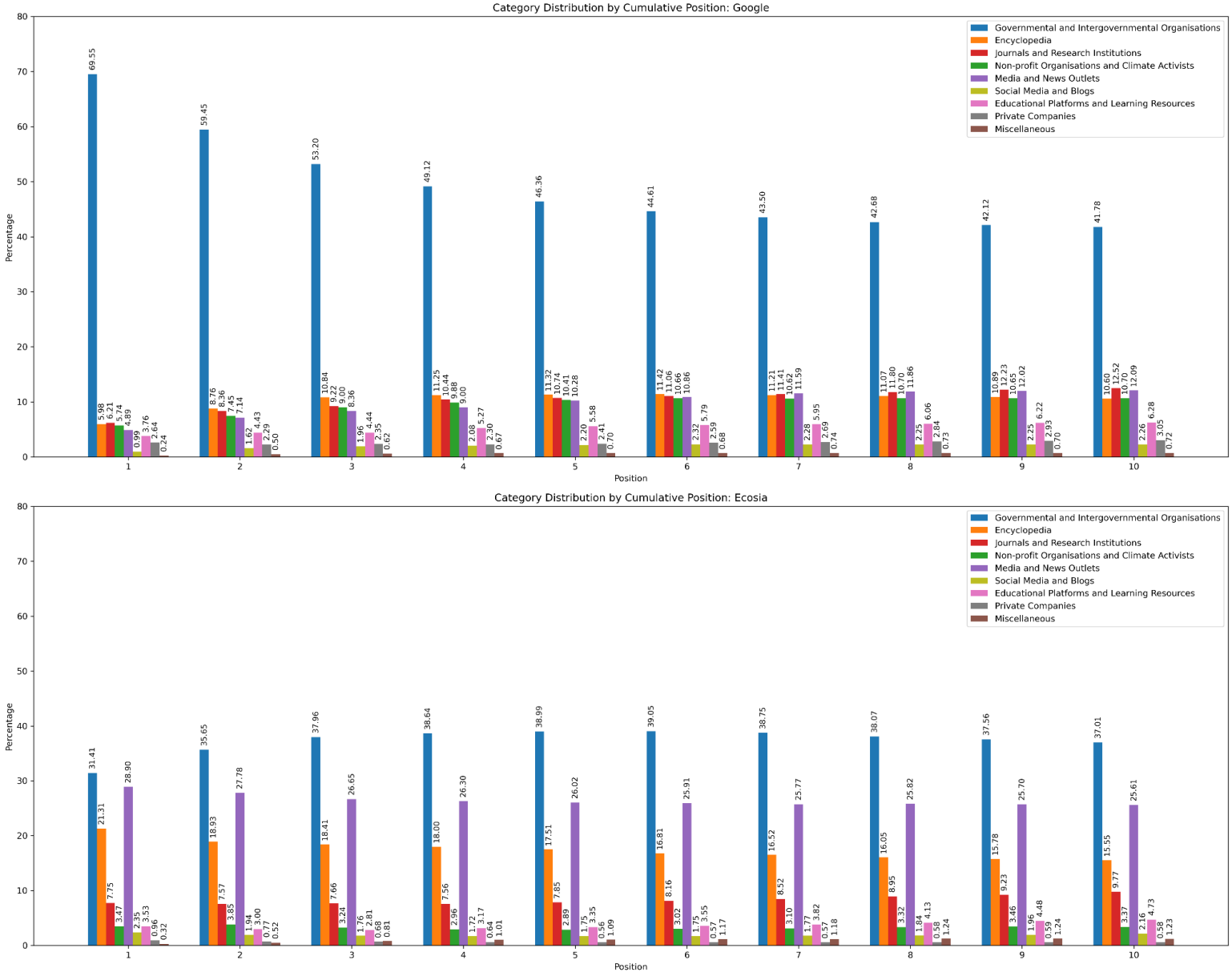
The domain appearing on Google in the first position the maximum number of times was ‘climate.nasa.gov’ with 289 appearances, closely followed by ‘epa.gov’ with 255. Both of these, in addition to the other top domains, fall under the category of “ Governmental and Intergovernmental Organisations”. The only exception to this was Wikipedia, which had the fifth highest occurrences in the first position at 95 and was classified as an “Encyclopedia”.

For Ecosia, the most frequent first position holder was National Geographic, with 330 appearances and was classified as “Media and Outlet”. Other top first position domains included NASA with 210 appearances which fell under “Governmental and Intergovernmental Organisations”, and Britannica and Wikipedia which fell under “Encyclopedia” with 206 and 158 appearances respectively.

Classification of Domains

The created categories of domains were utilised to determine the distribution of domain types across various positions for both search engines. As depicted in Figure 1, "Governmental and Intergovernmental Organisations" was the most frequently displayed category of domains by both Google and Ecosia through all positions.

Figure 1 *Percentage distribution of categories of displayed domains for cumulative data of all positions for Google and Ecosia.*



With 1478 occurrences out of 2390, constituting 68.9% of the share for the first position, this category dominated the Google search results significantly more than other categories (see figure 1). Although the numbers decreased from the second position onward, with 59.45% and 53.20% up to the second and third positions, respectively, the pattern was repeated for all positions, with 41.78% by the tenth position, and this category maintained its predominance.

"Encyclopedia" had a mere 5.98% share at the first position, which gradually increased to 8.76% and 10.84% by the second and third positions, respectively. The appearance of this category of domains showed only a slight increase, with 11.42% up to the sixth position, before experiencing a decline and reaching 10.60% by the tenth position. "Journals and Research Institutes" and "Media and News Outlets" shared similar patterns,with 6.21% and 4.89% shares at the first position, respectively. The numbers exhibited a gradual but consistent upward trend, with 10.44% and 9.00% by the fifth position, and 12.52% and 12.09% by the tenth position, respectively.

"Non Profit Organisations and Climate Activists" started off with 5.74% at the first position and, with a gradual increment, accumulated 10.41% shares by the fifth position. However, following this position, the shares plateaued and did not move significantly, ending with 10.70% by the tenth position. "Social Media and Blogs" and "Private Companies" were rarely displayed, with 0.99% and 2.64% shares at the first position, with very small increments with the decreasing ranks. The visibility of the former peaked up to the sixth position with 2.32% shares. The distribution of "Educational Platforms and Learning Resources" did not vary significantly, with 3.76% at the first position and a minimal increment with the added freedom of positions. It had 5.58% shares by the fifth position and 6.28% by the tenth. The "Miscellaneous" category, although low, contains only one domain of the World Economic

Forum. For the first position, it had 0.24% share, which increased to 0.70% by the fifth and 0.73% by the eighth.

Ecosia displayed a similar result when it came to the category of "Governmental and Intergovernmental Organizations" and it was the most frequently displayed category across all positions, however the shares, compared to Google, differed significantly . The category had 31.41% shares at the first position, which was much lower than that of Google. There was an increase in the appearance by the second position with 35.65%, after which it had an even visibility.

The second most popular category for Ecosia was "Media and News Outlets" with a contribution of 28.90% in first position. Its share remained fairly stable, with 26.02% by the fifth, 25.61% by the tenth, and similar numbers in all the intermediate positions. "Journals and Research Institutions" followed a similar stability trend, with 7.75% at the first position with minimal increment in visibility, ending with 9.77% by the tenth position. The exposure for the third most popular category, "Encyclopedia" was observed to peak at the first position with 21.31%, after which a steady drop was observed, and it was seen 17.51% of the time by the fifth position and 15.55% by the tenth. The share of "Non Profit Organisations and Climate Activists" for Ecosia was lower than that of Google with 3.47% at the first position, which dropped to 2.89% by the fifth position. A slight increase was observed with a 3.37% share by the tenth position. "Educational Platforms and Learning Resources" , "Social Media and Blogs" and "Private Companies" witnessed minimal exposure and increment. The "Miscellaneous" category, that is, World Economic Forum had a similar first position share as Google with 0.32%. The increment was higher than that of Google with 1.09% by the fifth position and 1.24% by the tenth.

Common Domain

By Count

Google and Ecosia had a total of 619 common domains, which accounted for 25.32% of the total unique domains in the data set. However, the frequency of some of these common domains' appearances differed vastly in Google and Ecosia. Britanicca, which appeared 1901 times in Ecosia, appeared only 659 times in Google. National Geographic and NASA exhibited an identical trend, with 1419 and 1830 instances for Ecosia, compared to 506 and 830 instances for Google, respectively. Some common domains had comparable instances, which included EPA, Wikipedia, and the UN, with 1119, 958, and 846 for Ecosia and 1056, 1038, and 580 for Google, respectively.

A category based comparison was also performed for the common domains, as illustrated in Table 5.

Table 5

Number of domains and percentage share for categories for domains that have been classified.

Category	Domain Count	Percentage Share
Governmental and Intergovernmental Organizations	73	26.94
Journals and Research Institutions	56	20.66
Media and News Outlets	51	18.82
Encyclopedia	25	9.23

Category	Domain Count	Percentage Share
Educational Platforms and Learning Resources	23	8.49
Non-profit Organisations and Climate Activists	20	7.38
Social Media and Blogs	14	5.17
Private Companies	8	2.95
Miscellaneous	1	0.37
	271	

Note. The remaining 348 domains are unclassified due to their relatively lower occurrences.

It was found that 73, out of the 271 classified domains, belonged to the "Governmental and Intergovernmental Organizations" category. This accounted for a 26.94% share. 56 domains from "Journals and Research Institutions" accounted for 20.66% of the share, while 51 domains from "Media and News Outlets" accounted for 18.82%.

Comparing the similarity of the results from Google and Ecosia with the Jaccard similarity index, as indicated in Table 6, yields an overall overlap of 25%, that is, when all ten positions are considered. The overlap for the first position was found to be 22%, while it was 24%, 25% and 26% for top two, three and five positions.

Table 6

Jaccard similarity index for domains of Google and Ecosia

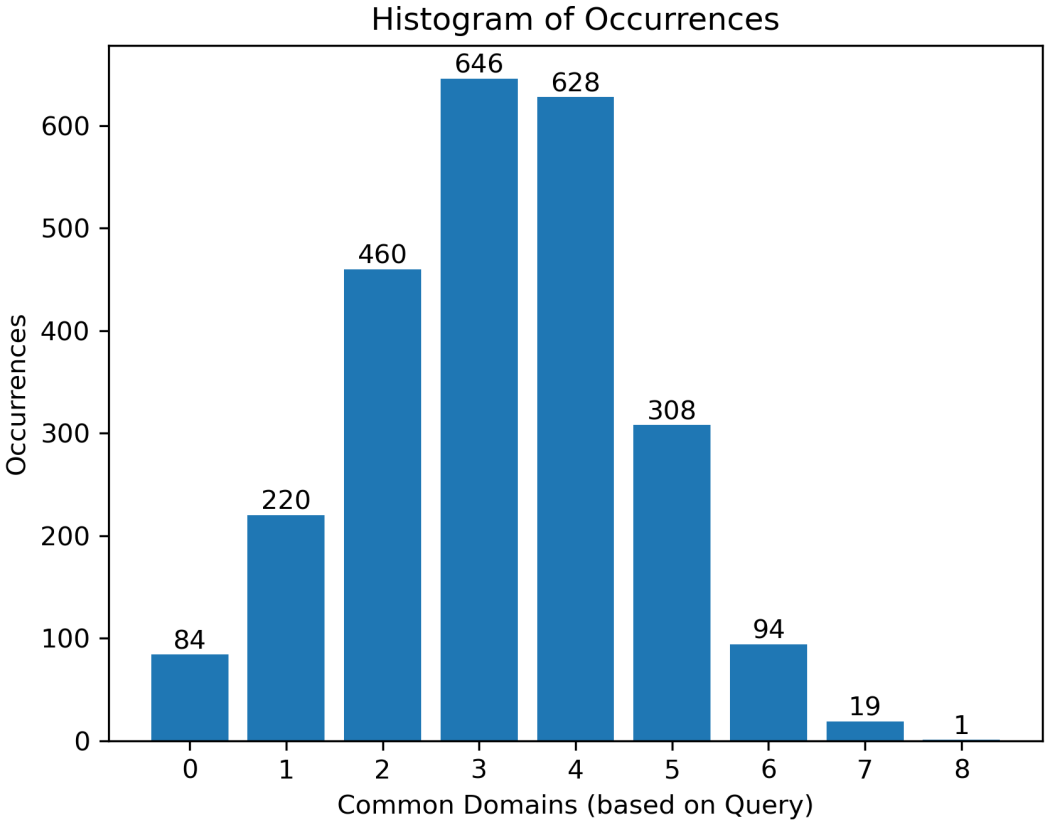
	Top 1	Top 2	Top 3	Top 5	Top 10
Jaccard Index	0.22	0.24	0.25	0.26	0.25

By Query

On evaluating the domains displayed by both search engines for every query and identifying the number of identical domains, numbers between zero and eight emerged (see figure 2).

Figure 2

Count of common domains match, out of 10, for Ecosia and Google for a given search query and the occurrences of the specific number of common domains.



The maximum number of common domains observed was eight, but this was only the case for one query, which was "short note on greenhouse effect". The highest number of queries, 646, had three common domains, closely followed by 628 queries with four common domains. For 84 search queries, no identical domains were found. The average number of common domains was found to be 3.18, while the median was 3.00.

Exclusive Domains

Google was found to have 1315 exclusive domains, while Ecosia had 509. The domain, 'myclimate.org' was the most frequently displayed Google-exclusive domain with 255 instances, followed by 'icos-cp.eu' with 96 instances (see table 7).

Table 7

Domains that are exclusive to Google and Ecosia separately, with the number of their occurrences.

Google		Ecosia	
Domain	Count	Domain	Count
www.myclimate.org	255	docs.google.com	74
www.icos-cp.eu	96	beta.wmo.int	70
www.climate.gov	72	hyperphysics.gsu.edu	49
climatechange.lta.org	65	timeforchange.org	21
www.tn.gov	64	www.msn.com	21
homework.study.com	63	www.ehow.com	18
sustainabilitymag.com	60	lamont.columbia.edu	18
www.climatechange.environment.nsw.gov.au	55	homeguides.sfgate.com	18
atmosphere.copernicus.eu	54	www.nas.nasa.gov	18
earthathome.org	51	egusphere.copernicus.org	17

Bing had previously shown up as Ecosia's most popular exclusive domain, with a staggering 756 instances, far more than the others. However, this was later found to be an error in the dataset, rooting from the fact that Ecosia sources its results from Bing, and was hence removed. Ecosia-exclusive domains included Google Docs, WMO, and GSU, albeit with much fewer instances.

Discussion and Conclusion

The objective of this study was to compare climate-related search results from Google and Ecosia. Wikipedia has typically been observed to be the most popular domain across all search engines for general queries and trends in previous studies (Yagci et al., 2022), yet it was surprisingly not the most popular in either of these two search engines for searches related to climate change. With 1038 occurrences, it was at the second position on Google, 18 behind EPA, while in Ecosia it was much lower, taking the sixth position with 958 occurrences. However, for both search engines, Wikipedia showed up in the first position several times, with a frequency of 95 and 158 for Google and Ecosia, respectively. It is also worth noting that nationalgeographic.com (1419) and nationalgeographic.org (1132) were counted as separate domains, and when combined, the total number of occurrences for National Geographic clearly surpass the other media with 2551 appearances for Ecosia. The top three most popular domains of Ecosia had close to twice the frequency compared to the top three of Google, hinting towards Ecosia being more repetitive with its search results. The fact that Ecosia has 1127 unique domains, almost half as many as Google, which has 1934, lends support to this hypothesis even further.

As a measure of dispersion, the standard deviation was used (Pérez-Iglesias & Araujo, 2010). Based on the position of the results, both search engines displayed a standard deviation between 2-3 for most of their domains, indicating an overall moderate dispersion of results. When arranged in ascending order of standard deviation, only 100 of the bottom domains for Ecosia had standard deviation values greater than the specified range, whereas for Google it was 200 of the bottom domains. This result is consistent with the fact that Google has close to twice as many domains as Ecosia.

Google and Ecosia shared 619 common domains, with over one quarter (26.94%) falling under the category of “Governmental and Intergovernmental Organisations”. This finding suggests that these organisations play a significant role in providing information for climate-themed queries. As an extension, it indicates a strong presence of governmental and intergovernmental organisations in addressing climate-related issues. Even in the individual evaluation of the search engines, this category exhibited popularity on both Google and Ecosia. However, it was clear that Google yielded a significantly higher share of results within this category.

The other two major categories were “Journals and Research Institutes”, and “Media and News Outlets” , suggesting that for climate-themed searches, trustworthy and verified generic sources are deemed more relevant by both search engines compared to websites of nonprofit organisations, personal blogs and social media. Overall, Ecosia showed a broader spectrum of categories across all ranks, demonstrating a more diverse range of domains, and indicating a comprehensive and diverse set of sources. In contrast, Google predominantly displayed domains from the category of "Governmental and Intergovernmental Organisations.", suggesting a strong inclination towards official sources for results.

When comparing the common domains on the basis of results displayed by individual queries, it was found that the histogram was concentrated between 3 to 4 common domains per query. The mean and the median of the data were quite close, with values of 3.18 and 3.00 respectively, suggesting that there are no significant abnormalities affecting the results in this section.

With the aid of the Jaccard Index, the maximum overlap between the two search engines was found to be 26% for the top five results, while for the entire dataset of the top 10 results, the overlap was 25%. This finding is consistent with earlier research by Makhortykh et al. in

2020, which shows that Google and Bing exhibit a low level of result overlap. Given that Ecosia sources its results from Microsoft Bing, the results obtained in this study are relevant and applicable to the context of this investigation.

A limitation of the study was that the employed data set was retrieved from a single location, thereby possibly not capturing the variations that could have resulted from analysing a data set from a broader geographical region. The influence of location on search engine results is an essential factor to consider. Future studies with a broader geographic scope would provide a more in-depth understanding of the broader implications and potential variations in the results from Google and Ecosia across different regions.

Another limitation of the study was the focus on only three broad climate-related themes for query generation. This narrow focus may not cover all of the queries related to climate-related topics that users may want to make. Restricting the query generation could lead to potentially overlooking other important aspects and results. Future studies could benefit from incorporating a more comprehensive set of themes or keywords to ensure a more thorough analysis of the behaviour of search engines for climate-related queries.

Lastly, an interesting area to explore in future research would be the possibility that two distinct search engines could yield completely different sets of results while both remaining comparably relevant. Looking into these open questions could potentially illuminate search engine functionality and inform future search experiences.

References

Allianz | LOHAS anyone? (2009, February 11). Allianz.com. Retrieved June 6, 2023, from https://www.allianz.com/en/press/news/studies/news-2009-02-11.html

Bar-Ilan, J., Mat-Hassan, M., & Levene, M. (2006, July). Methods for comparing rankings of search engine results. *Computer Networks*, *50*(10), 1448–1463. https://doi.org/10.1016/j.comnet.2005.10.020

Behnert, C., & Lewandowski, D. (2015, November). Ranking Search Results in Library Information Systems — Considering Ranking Approaches Adapted From Web Search Engines. *The Journal of Academic Librarianship*, *41*(6), 725–735. https://doi.org/10.1016/j.acalib.2015.07.010

Chiou, L., & Tucker, C. (2017). Search Engines and Data Retention: Implications for Privacy and Antitrust. *National Bureau of Economic Research*. https://doi.org/10.3386/w23815

Choi, D., Gao, Z., & Jiang, W. (2020, February 14). Attention to Global Warming. *The Review of Financial Studies*, 33(3), 1112–1145. https://doi.org/10.1093/rfs/hhz086

Choi, S., & Feinberg, R. A. (2021, February 3). The LOHAS (Lifestyle of Health and Sustainability) Scale Development and Validation. *Sustainability*, *13*(4), 1598. https://doi.org/10.3390/su13041598

Ecosia GmbH. (2023). Retrieved May 7, 2023, from https://www.ecosia.org

Enge, Spencer, & Stricchiola. (2015, August 17). *The Art of SEO: Mastering Search Engine Optimization* (3rd ed.). O'Reilly Media, Inc.

Erdmann, A., Arilla, R., & Ponzoa, J. M. (2022, May). Search engine optimization: The long-term strategy of keyword choice. *Journal of Business Research*, *144*, 650–662. https://doi.org/10.1016/j.jbusres.2022.01.065

Gezici, G., Lipani, A., Saygin, Y., & Yilmaz, E. (2021, January 27). Evaluation metrics for measuring bias in search engine results. *Information Retrieval Journal*, *24*(2), 85–113. https://doi.org/10.1007/s10791-020-09386-w

Guan, Z., & Cutrell, E. (2007, April 29). An eye tracking study of the effect of target rank on web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/1240624.1240691

Hinman. (2018, August 31). *Searching Ethics: Ethical Issues in Search Engine Technology – The Research Center on Values in Emerging Science and Technology*. Searching Ethics: Ethical Issues in Search Engine Technology – the Research Center on Values in Emerging Science and Technology. Retrieved May 18, 2023, from https://rcvest.southernct.edu/searching-ethics-ethical-issues-in-search-engine-technology/

Jungblut, & Stolz. (2019, December 30). *Our Digital Carbon Footprint: What's the Environmental Impact of the Online World?* Digital for Good | RESET.ORG. Retrieved June 6, 2023, from https://en.reset.org/our-digital-carbon-footprint-environmental-impact-living-life-online-12272019/

Krupp, B., & Gersey, J. (2022, December). Privacy Focused Companies, How Focused Are They? *ACM SIGCAS Computers and Society*, *51*(3), 10–10. https://doi.org/10.1145/3585060.3585064

Lewandowski, D. (2015, February 23). Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, *66*(9), 1763–1775. https://doi.org/10.1002/asi.23304

Lewandowski, D. (2017). Is Google Responsible for Providing Fair and Unbiased Results? *Law, Governance and Technology Series*, 61–77. https://doi.org/10.1007/978-3-319-47852-4_4

Lewandowski, D. (2021, June 29). *Suchmaschinen Verstehen*. Springer Vieweg. https://doi.org/10.1007/978-3-662-63191-1

Lewandowski, D., & Sünkler, S. (2013, June 6). Designing search engine retrieval effectiveness tests with RAT. *Information Services & Use*, *33*(1), 53–59. https://doi.org/10.3233/isu-130691

Lewandowski, D., & Sünkler, S. (2019, January 31). Das Relevance Assessment Tool. *Information - Wissenschaft & Praxis*, *70*(1), 46–56. https://doi.org/10.1515/iwp-2019-0007

Maillé, P., Maudet, G., Simon, M., & Tuffin, B. (2022, February). Are Search Engines Biased? Detecting and Reducing Bias using Meta Search Engines. *Electronic Commerce Research and Applications*, 101132. https://doi.org/10.1016/j.elerap.2022.101132

Makhortykh, M., Urman, A., & Ulloa, R. (2020, May 11). How search engines disseminate information about COVID-19 and why they should do

better. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-017

Parsania, Kalyani, & Kamani. (2016, December). A Comparative Analysis: DuckDuckGo Vs. Google Search Engine. *GRD Journal for Engineering*, *2*(1). https://grdjournals.com/uploads/article/GRDJE/V02/I01/0002/GRDJEV02I010002.pdf

Pérez-Iglesias, & Araujo . (2010). Standard Deviation as a Query Hardness Estimator. *Chavez, E., Lonardi, S. (Eds) String Processing and Information Retrieval. SPIRE 2010. Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-642-16321-0_21

Rosenkrantz, W. A. (2008, July 10). *Introduction to Probability and Statistics for Science, Engineering, and Finance*. Chapman & Hall/CRC. https://doi.org/10.1604/9781584888123

Ruch, Schmidt, Decker, & Kolbe. (2011). Ecosia–Who Cares About a Green Search Engine? . *AMCIS 2011 Proceedings - All Submissions. 304.* . https://aisel.aisnet.org/amcis2011_submissions/304

Schmidt, N. H. (2011, January). Search Engines and Social Business – Implications from the Case of Ecosia. *Environmentally Sustainable Information Management: Theories and Concepts for Sustainability, Green IS, and Green IT* .

Segev, E. (2010). *Google and the Digital Divide: The Bias of Online Knowledge*. Chandos Publishing.

Sharma, P., & Dash, B. (2022, June 30). The Digital Carbon Footprint: Threat to an Environmentally Sustainable Future. *International Journal of Computer*

Science and Information Technology, *14*(03), 19–29. https://doi.org/10.5121/ijcsit.2022.14302

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999, September). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, *33*(1), 6–12. https://doi.org/10.1145/331403.331405

Sirisuriya , D. S. (2015). A comparative study on web scraping. *Proceedings of 8th International Research Conference, KDU*.

Sisco, M. R., Pianta, S., Weber, E. U., & Bosetti, V. (2021, June). Global climate marches sharply raise attention to climate change: Analysis of climate search behavior in 46 countries. *Journal of Environmental Psychology*, *75*, 101596. https://doi.org/10.1016/j.jenvp.2021.101596

Statcounter. (2023). StatCounter Global Stats. Retrieved June 6, 2023, from https://gs.statcounter.com/search-engine-market-share

UNEP. (2021, March 19). UNEP. Retrieved June 6, 2023, from http://www.unep.org/news-and-stories/story/new-pact-tech-companies-take-climate-change

Wang, X., & Pleimling, M. (2017, March 29). Foraging patterns in online searches. *Physical Review E*, *95*(3). https://doi.org/10.1103/physreve.95.032145

Wellings, S., & Casselden, B. (2019, November 26). An exploration into the information-seeking behaviours of engineers and scientists. *Journal of Librarianship and Information Science*, *51*(3), 789–800. https://doi.org/10.1177/0961000617742466

Yagci, N., Sünkler, S., Häußler, H., & Lewandowski, D. (2022, October). A Comparison of Source Distribution and Result Overlap in Web Search

Engines. *Proceedings of the Association for Information Science and Technology*, *59*(1), 346–357. https://doi.org/10.1002/pra2.758