Measuring Consensus and Disagreement Between Search Engines

Felix Park
University of Virginia
fjp7mb@virginia.edu

Andrea Chang
University of Virginia
ayc3ue@virginia.edu

Aditi Narvekar
University of Virginia
an5vq@virginia.edu

Nikhil Bhaip
University of Virginia
nb5hd@virginia.edu

ABSTRACT

Search engines strive to return the most likely relevant documents from queries, but relevance is very subjective. Commercial search engines like Google, Bing, DuckDuckGo, and Baidu all have the goal of returning relevant documents, but return their own distinct results for the same queries. In this work, we evaluate the consensus and disagreement between the results of various search engines and different topics across search engines. We use various consensus measures for the evaluation process including Rank Biased Overlap, proposed by Weber et al, and Consensus Ranking. We also propose an altered version of AnchorMAP, originally introduced by Buckley, in order to offer a symmetric metric for calculating similarities between rankings.

The results of our evaluation metrics indicate that Bing and Baidu differ the most, and Bing and DuckDuckGo are the most similar. This interesting observation is shown to be statistically significant through the Wilcoxon Signed-Rank Test.

CCS CONCEPTS

• **Information Systems → Information Retrieval**.

KEYWORDS

ACM Proceedings, Information Retrieval, Rank Correlation, Consensus Ranking, Google, Baidu, Bing, DuckDuckGo

ACM Reference Format:
Felix Park, Andrea Chang, Aditi Narvekar, and Nikhil Bhaip. 2018. Measuring Consensus and Disagreement Between Search Engines. In *Proceedings of Dec 2018 (UVA IR).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

The goal of a search engine is to return the most relevant documents from a given query; however, the definition of relevance seems to vary among people, research groups, and companies. Different commercial search engines offer unique orderings of results or even completely distinct results for the same query. Additionally, differences between search engines may be more pronounced for certain topics, especially between topics that reflect a cultural or political difference.

The goal of this project is to compare and contrast search result rankings across various search engines. The search engines of interest include popular ones like Google and Bing, as well as the less frequently used DuckDuckGo, which is known for not tracking user behavior. Additionally, we will compare with a foreign search engine, Baidu, which is specifically known for censorship of politically controversial topics. Comparing to a foreign search engine may also reveal cultural or political differences that are worthy of investigation.

We conducted a formal exploratory data analysis into the ranked results in the aforementioned search engines. We hope these results can tell us the similarities and disparities between search engines and whether these results differ across various subjects and domains. Furthermore, we used several metrics for measuring consensus, some of which are novel approaches proposed in the research community.

2 BACKGROUND

There are several metrics we use to compare the results from the various search engines. In this section, we will briefly discuss some of the popular metrics. In the Related Works section, we discuss more modern approaches to calculating consensus introduced in the scientific community.

2.1 Jaccard Correlation

Jaccard correlation is a straightforward statistic to measure the similarity between two sets and is defined as the cardinality of the intersection divided by the cardinality of the union:

$$J(L_1, L_2) = \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}$$

The closer the calculated Jaccard similarity value is to 1, the more similar the two sets are. This is a fairly naive approach because the Jaccard similarity value does not weight all of the documents equally regardless of rank. If the documents are disjoint (have no documents in the intersection), then the Jaccard value is 0.

UVA IR, Charlottesville, VA

ACM ISBN 123-4567-89-123/08/06…$15.00
https://doi.org/10.475/123_4

2.2 Top K Symmetric Difference

Another approach to comparing two ranked sets is considering what the sets don't have in common, using a measure called symmetric difference. Symmetric difference is defined as the intersection of two sets subtracted from their union. It can be thought of as an XOR operation. For the purposes of this work, we use the cardinality of the symmetric difference and normalize it by dividing by the sum of the size of the two lists. However, symmetric difference alone is inadequate. For example, if two ranked results of 10 items were exact reverses of each other, the symmetric difference would indicate that they were similar rankings.

Therefore, we use a metric called top K symmetric difference, proposed by Fagin, et. al. [5], which calculates the symmetric difference for the top K items at every point from 1 through K. The formula for top K symmetric difference is as follows:

$$SD(L_1, L_2) = \sum_{n=1}^{k} \frac{1}{2k}(|L_1 \cup L_2| - |L_1 \cap L_2|)$$

2.3 Rank-Biased Overlap for Indefinite Rankings

Search engine results may be highly varied. The depth, or number of results for a query, may differ between search engines, although we only examine the top 5 results. The results between search engines may also be disjoint (have no intersection), or may have some intersection, but differ in relative rankings. Rank-biased overlap (RBO) [11] is a similarity metric that weights similarities relative to depth, giving greater weight to similarities observed at higher rankings.

RBO's weighting is also bounded, so the potentially infinite tail set of documents are never weighted more than the finite head set of documents. The dataset used in this paper only examines the top 5 results, so this is not a concern, but is a consideration that RBO has. RBO also can be used for disjoint sets, giving a measure for similarity of the full ranking, and then estimating a similarity value based on the truncated list of results.

For two lists of results L_1 and L_2, the intersection of L_1 and L_2 at depth d is denoted I. The weighting p denotes the probability the user would examine the next document, deriving its value from a geometric progression of depth d. The geometric progression formula, shown below, puts an upper bound on p so that $0 < p < 1$.

$$\sum_{d=1}^{\infty}(p^{d-1}) = \frac{1}{1-p}$$

The resulting RBO formula, combining these elements, is shown below:

$$RBO(L_1, L_2, p) = (1-p)\sum_{d=1}^{\infty}(p^{d-1})\frac{|I|}{d}$$

This RBO metric needs an infinite ranking of lists L_1 and L_2, but a reasonable extrapolation can be made to approximate a single value. A variant of RBO was used to calculate the similarity.

2.4 AnchorMAP

AnchorMAP, proposed by Buckley [3], is another rank similarity metric that compares two ranked lists by assuming that the top k documents from one list are relevant and then computing MAP for the other list. AnchorMAP is a simple metric, but is shown to have good performance.

Something we noticed was that AnchorMAP was not symmetric. How could we decide which list to use as the "relevant" list? In order to introduce symmetry, we propose an altered AnchorMAP which involves calculating the AnchorMAP twice and taking the average. Each time, we assume one of the lists is the "relevant" list.

$$Anchor'(L_1, L_2) = \frac{Anchor(L_1, L_2) + Anchor(L_2, L_1)}{2}$$

This altered metric was motivated by Yilmaz et al.'s work on introducing a symmetric AP correlation metric based on Kendall's Tau metric [12].

3 RELATED WORKS

In this section, we will go over more modern works related to measuring consensus between search engines.

3.1 Comparing Google Maps and Bing Maps

In Cipeluch et al.'s work, the authors analyzed the accuracy of Google Maps and Bing Maps based on spatial coverage, currency, and positional accuracy [4]. While there was no clear winner, they found that each platform showed individual differences and similarities for each case.

3.2 Consensus Measure of Rankings

In *Consensus measure of rankings* by Lin et al. [6], the authors evaluate the degrees to which the rankings between search engines agree. These similarity measures are useful to evaluate consensus when the ground truth is not truly available for evaluation. In their work, the authors introduce an approach for consensus measures using graph representations.

3.3 Consensus Ranking

It is difficult to label the ground truth for rankings. Ground truth is defined as the ideal ordering of relevant results. While the goal of this project is not to determine ground truth, a consensus ranking of two ranked lists, denoted L_C, can act as a ground truth and help determine the agreement of the two rankings. The consensus rank algorithm used in this experiment finds a ranking that agrees the most with 2 given ranking sets.

Meila et al. proposed some recent work on estimating a consensus ranking based on search methods on the Mallows Model [7]. we need to say some shit on this if we can make sense of it at all.

4 METHODOLOGY

4.1 Data collection

We constructed a dataset of queries for our experiments; however, we also chose to separate them by categories in order to analyze the similarities and differences across different topics as well. The

various subjects included generic words, religion, current events, US history, abstract ideas, and science.

4.1.1 Generic words. The generic terms were sourced from a Kaggle dataset [10] which contained the 1 million most common words in the English language. Stop words like "the" and "an" were ignored, and only specific terms like "companies" and "delivery" were included as part of the set. We randomly selected 500 general terms in this set. Our hope was that this dataset would indicate how these search engines treat neutral topics.

4.1.2 Religion. The religion dataset consists of 100 terms such as "Buddha", "Vishnu", and "Church". The terms were selected randomly across different religions from an online religion dictionary [1]. We hypothesized the religious terms would reveal cultural differences in information need for each search engine's users.

4.1.3 Current events. 100 current events queries were taken from Google Search trends. This meant that this dataset was already heavily biased towards Google, since they came from trending topics on Google. However, current events change day to day, and curated datasets were not available, so this seemed like the best alternative. Because current events queries commonly return news sources, we hypothesized that the results would reveal each search engine's dependence on news sources.

4.1.4 US History. There were 100 US History terms were taken from Quizlet study flashcards. Queries in the history list included "American Revolution" and "The Boston Tea Party" [9]. The purpose of including US History terms in the dataset was to reveal differences in topics sourced from an education environment.

4.1.5 Abstract Ideas. We constructed a dataset of abstract ideas with terms like "dedication" and "mercy". The terms in this set might not have a formal definition, and we were curious to see if there were varying results here. We also hypothesized that abstract ideas could reveal cultural differences in their interpretation.

4.1.6 Science. With short queries like those in abstract ideas and generic words, we were concerned that the results might lead to dictionary definitions or Wikipedia pages and would not provide interesting results. Thus, we constructed the science dataset to be a list of open ended questions that might not have a straightforward answer. These question queries included things like "how does bm25 work", "how is cancer treated today", and "what is mendel famous for". We hoped that by representing queries as questions, there might be variance in different search engine results, revealing if formulating a query in a question format provided a greater variety of results.

4.2 Scraping search engines

We queried the search engines Google, Bing, Baidu, and DuckDuckGo using the aforementioned query datasets. We collected all of the search results on the first page, excluding the results that were not in list form, such as Google's Knowledge Graphs, Google's Rich Answers, and advertisements. We chose to scrape the first page because these were the results users were more likely to see. However, each query returned varying quantities of results across the search engines. We chose to keep the top 5 results as this was the minimum quantity all search engines returned and users were likely to see the first 5 results.

We scraped the results for DuckDuckGo, Bing, and Baidu using a open-source web scraper called GoogleScraper [8], and our own Python script for scraping Google using ScraperAPI.com [2]. GoogleScraper bypasses bot-detecting efforts that search engines implement, allowing automation of a large number of queries into each search engine without getting blocked. To bypass bot detection, GoogleScraper uses the browser emulator Selenium for each query, allowing us to query as if the query originated from its own machine. The ScraperAPI also has its own measures for avoiding bot detection.

4.3 Result Normalization

For each query, the results that were collected were in the form of URLs, most of which contained extraneous characters at the end of the URL that were irrelevant to identifying the content of the document. To exclude irrelevant portions of the URL, we normalized all URLs to include only the website name and the domain name. For example, a website would be normalized from "https://www.history.com/topics/british-history/enlightenment" to "history.com". This normalization procedure also accounts for small differences in the URLs between search engines, allowing us to assume that a result was identical if only the website name and domain matched.

4.4 Implementing Evaluation Metrics

We implemented Jaccard correlation, Top k Symmetric Difference, RBO, and Anchormap on the normalized search results for all above queries. We then implemented a consensus ranking algorithm to find the distances between consensus ranking based on an objective function as well as the correlation between the original rankings and the consensus ranking.

Our approach to implementing consensus rank involved listing all combinations of elements in the union of L_1 and L_2, permuting every combination in the list of all combinations, and calculating the similarity between the consensus rank and L_1 and L_2. For example, if lists L_1 and L_2 each contained 5 results, we combine L_1 and L_2 into a single list of length 10. Permutations of length 5 would then be taken from this combined listing, and the permutation with the highest similarity value would be considered the consensus ranking.

5 RESULTS

We performed a significant amount of analysis on the data we collected and show the most interesting ones in this section. The rest of the analysis can be found in the Appendix.

5.1 Distribution

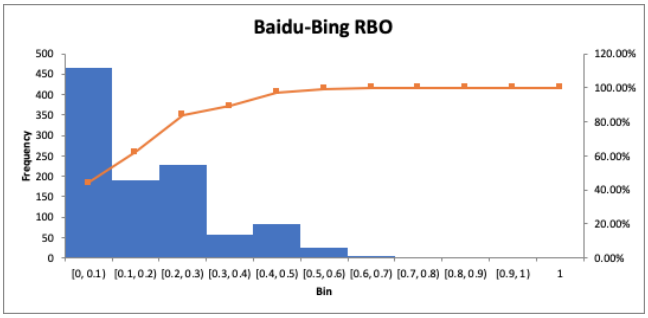


Figure 1: This histogram shows the distribution of the RBO Metric between the rankings from Baidu and Bing

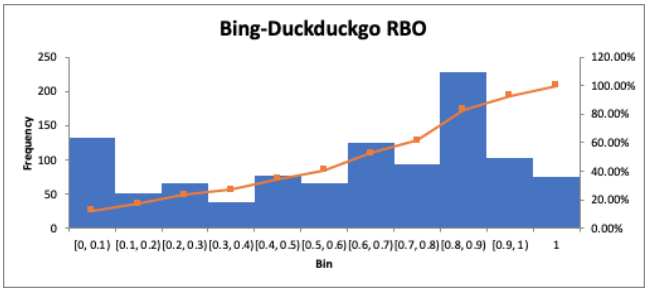


Figure 2: This histogram shows the distribution of the RBO Metric between the rankings from Bing and DuckDuckGoogle

In Figure 1, we notice that the RBO Similarity metric is mostly distributed towards 0. This gives us an impression that Baidu and Bing tend to disagree with the documents they return. In contrast, Figure 2 shows that Bing and DuckDuckGo return very similar rankings from each other.

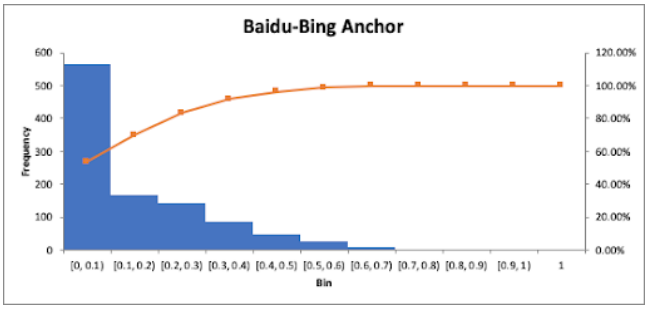


Figure 3: This histogram shows the distribution of the AnchorMAP Metric between the rankings from Baidu and Bing

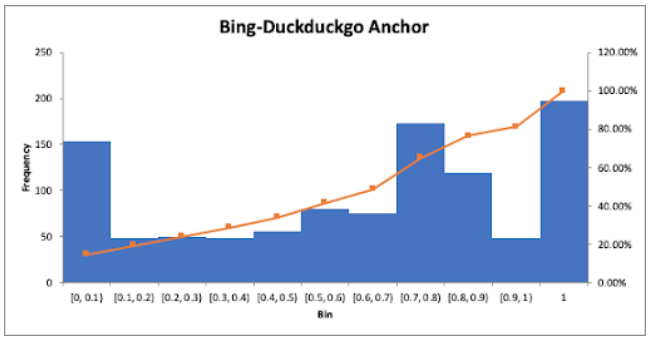


Figure 4: This histogram shows the distribution of the AnchorMAP Metric between the rankings from Bing and DuckDuckGoogle

In Figure 3, we observe the same results as what we saw from RBO that Baidu and Bing differ greatly. In Figure 4, we see more evidence that supports the RBO measures that Bing and DDG tend to have a consensus on their rankings.

5.2 Statistical Analysis

As shown in the results above, we couldn't perform a Pairwise Student t-test. The Student t-test assumes a normal distribution, which our data doesn't show. Thus a non-parametric signed test would fit the task better, so the Wilcoxon Signed-Rank Test was chosen for our statistical analysis.

AnchorMAP Statistics

	Baidu-Bing	Baidu-DDG	Baidu-Google	Bing-DDG	Bing-Google	DDG-Google
Median	0.075	0.083	0.120	0.710	0.367	0.487
Mean	0.129	0.133	0.145	0.591	0.369	0.469

Table 1: The table shows the AnchorMAP statistics for each of the search engine pairs

Baidu-Bing Wilcoxon Signed-Rank Test (Anchor)

	Baidu-DDG	Baidu-Google	Bing-DDG	Bing-Google	DDG-Google
p-value	7.056e-2	7.354e-6	4.089e-142	2.311e-120	1.011e-157

Table 2: The table shows the results for the Wilcoxon Signed-Rank Test between Baidu-Bing's AnchorMAP metric against the other search engine pairs

In Table 1, we notice that Bing-DuckDuckGo have a high correlation and Baidu-Bing do not. In Table 2, we see that all of the pairs are statistically significant except for Baidu-Bing v. Baidu-DuckDuckGo. It makes sense to reject the Null Hypothesis here because Bing and DuckDuckGo are highly correlated.

5.3 Language Bias in Search Engines

At first, we thought that Baidu differed greatly because Baidu is a Chinese company and represents a different culture. Although this was true, it was interesting to see that the correlation between Google-Bing and Google-DuckDuckGo was not nearly as great as Bing-DuckDuckGo.

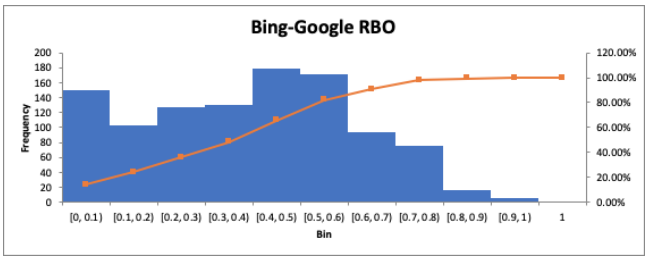


Figure 5: This histogram shows the distribution of the RBO Metric between the rankings from Bing and Google

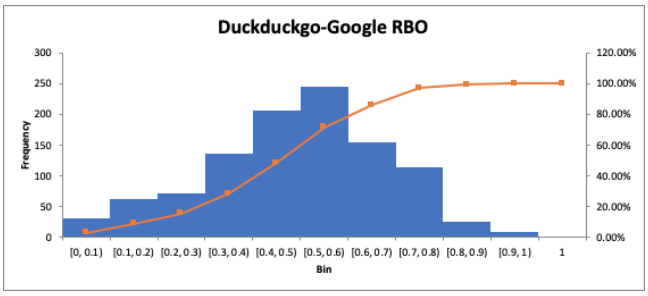


Figure 6: This histogram shows the distribution of the RBO Metric between the rankings from DuckDuckGo and Google

In Figure 5 and Figure 6, we see a relatively good RBO correlation, but it's nowhere near the correlation distribution from Bing and DuckDuckGo. This shows that the similarity between Bing and DuckDuckGo isn't solely due to the fact that it's an English search engine. It may be due to a much deeper reason on how these companies' algorithms determine document relevance.

Figure 8 and Figure 4, showing the AnchorMAP values, confirm these results as well.

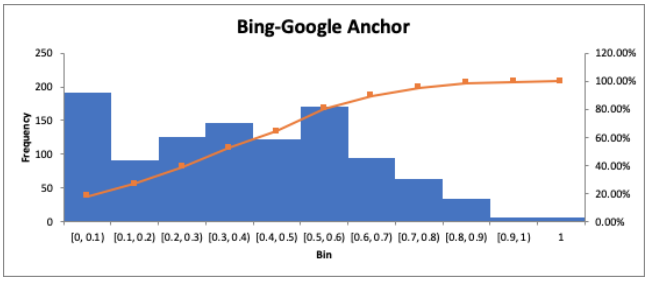


Figure 7: This histogram shows the distribution of the AnchorMAP Metric between the rankings from Bing and Google

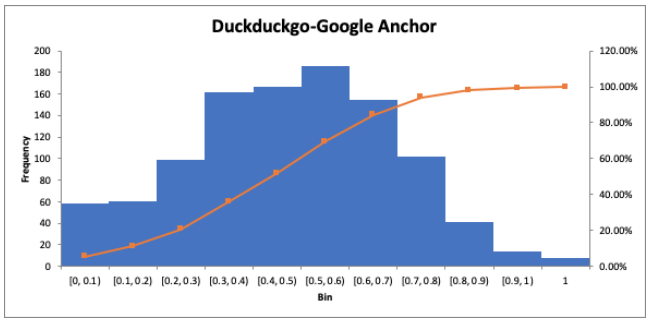


Figure 8: This histogram shows the distribution of the AnchorMAP Metric between the rankings from DuckDuckGo and Google

5.4 Exploring Consensus Across Topics

Here, we see results on how specific queries differ between Bing-DuckDuckGo and Baidu-Bing. As mentioned, queries are split by topics. In this section, we will go over the results of some of the more interesting topics. Most of the topics seemed to show the same results: Bing and DuckDuckGo agreed the most and Baidu-Bing agreed the least. However, for queries that were related to current events, Bing, DuckDuckGo, and Baidu all seemed to disagree very slightly.

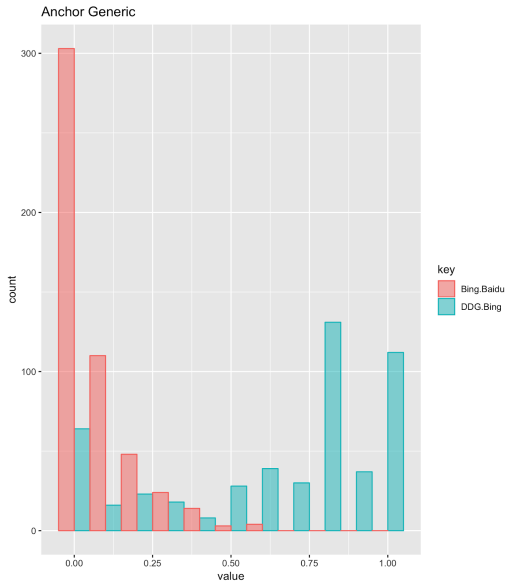


Figure 9: This histogram shows how correlations differ between Bing-DuckDuckGo and Baidu-Bing for the generic topics

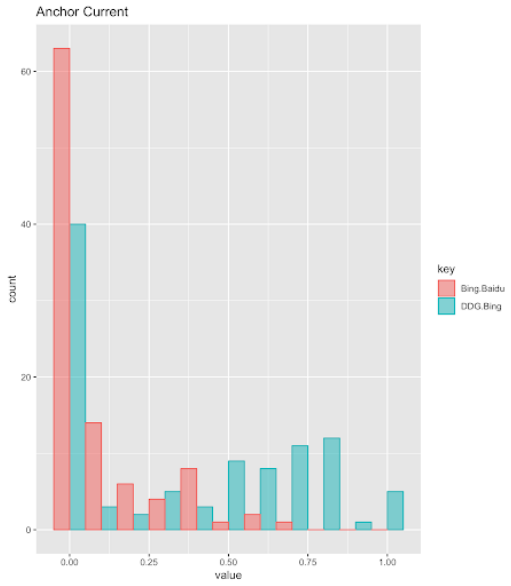


Figure 10: This histogram shows how correlations differ between Bing-DuckDuckGo and Baidu-Bing for the current events topics

Figure 9 illustrates how the two pairs are normally different through the other topics. However, in Figure 10, the two pairs both seem to disagree on their rankings. Although we observe this qualitatively, DuckDuckGo and Bing agree on a good chunk of the queries, giving it a bimodal-like distribution and pushing the p-value to be statistically significant. Overall, DuckDuckGo-Bing are similar and Bing-Baidu have different rankings across pretty much all topics. These results are backed by the Wilicoxon Signed-Rank Test in Table 3.

Bing-DDG vs. Baidu-Bing Wilcoxon Test

	Ideas	Current	Religion	Generic	History	Science
AnchorMAP	5.27055e-5	2.066879e-10	2.817507e-14	5.662525e-74	1.586832e-13	4.640156e-13
RBO	7.715448e-22	1.476124e-9	1.25737e-14	4.128162e-74	9.992932e-14	2.518737e-13

Table 3: The table shows p-values for the AnchorMAP and RBO metrics for Bing-DuckDuckGo similarity v. Baidu-Bing similarity

6 DISCUSSION

6.1 Bing, DuckDuckGo, and Baidu

From our results, we saw that Bing and DuckDuckGo returned very similar rankings while Bing and Baidu disagreed in their rankings according to the AnchorMAP and RBO results. The same relationship is seen in the jaccard correlation and top k symmetric difference, Appendix section A1 and A4. By taking a closer look at the rankings themselves, we can gain an intuition on why these were the results.

Firstly, it seems like Bing and DuckDuckGo assume one word queries are mostly definition queries. For example, Bing and DuckDuckGo would assume that the query "advertise" is asking for the definition of "advertise" while Baidu returned results pertaining to advertising on the web like "Yahoo Advertising" and "Advertising in WIRED". Another reason why Bing and Baidu differed greatly is because Baidu seemed to promote its own tools much more than returning other documents on the web. For the generic topic, Baidu returned nearly one Baidu service for each query! In contrast, we only saw 3 queries return Bing services from Bing (Bing Images and Bing Videos). The final and more obvious reason for this distinction is that Baidu offered more Chinese websites and domains than Bing and DuckDuckGo. This cultural difference makes sense since Baidu is a Chinese company.

It was also very interesting how Bing and DuckDuckGo showed a very large correlation that was much more different than Bing-Google and DuckDuckGo-Google. Google showed a similar trend to the first attribute we noticed from Baidu: the "advertise" query returned results related to advertising and news rather than definitions. It did return definitions, but it ranked the advertising platforms much higher. In contrast, Bing and DuckDuckGo both returned Merriam-Webster, dictionary.com, and thesaurus.com as their top 3 results in order.

6.2 Jaccard and Top K Symmetric Difference

The Jaccard and Top K Symmetric Difference metrics were more naive calculations, neglecting to consider a variety of parameters that influence the similarity value. The relative rankings, the depth of examination, and similarity for disjoint rankings all affected the similarity value, and were not parameters in the Jaccard correlation. Top K symmetric difference accounted for relative rankings and depth of examination, but did not directly directly give a greater weight to similarities at a higher rank. Top K symmetric difference also fails to account for differences in relative ranking as k increases in size.

For these reasons, we saw that the results from Jaccard and Top k SymDiff were not as smooth as the ones from RBO and AnchorMAP. There were very few unique values, which made the distributions look more discrete than continuous. Although they were naive, at the very least, they still confirmed what RBO and AnchorMAP revealed to a certain extent. due to the robustness of RBO and AnchorMAP, we focused majority of our analysis on those metrics; however, we did include results from Jaccard and Top K SymDiff in the Appendix.

6.3 Consensus Ranking

Although consensus ranking would be an interesting thing to explore, our results didn't reveal any helpful insights. First of all, our approach of creating the consensus ranking was very naive. In addition, we ran into bias issues when the two search engines gave disjoint results. The consensus ranking would usually take 3 results from one ranking and 2 results from another ranking; however, because the permutations were deterministic, keeping the consensus ranking with the maximum objective metric meant it would always keep the same 5 results based on position from the two rankings. This bias rendered our results useless and further analysis of our results would be uninteresting.

7 CONCLUSION

The objective of the experiment in this paper was to determine which search engines agreed and disagreed the most, and on which topics. We accomplished this goal by scraping Baidu, Bing, DuckDuckGo, and Google and comparing the ranked results with various evaluation metrics. These metrics included Jaccard correlation, Top K Symmetric Difference, Rank-Biased Overlap, and AnchorMAP.

Our results support the conclusion that Bing and DuckDuckGo have a statistically significant correlation. Bing and Baidu differ the most, and this difference is statistically significant as well based on Table 2. All four metrics support these conclusions. Jaccard and Top k SymDiff were naive metrics with very few unique values, which made the results look discrete compared to the more smooth curves of the other two metrics.

One might assume that Baidu is the most different from the search engines because it is a Chinese company, and therefore prioritizes different information than the American search engines. However, if we view Google as a control, we don't see a large correlation between Google and Bing or Google and DuckDuckGo. This might suggest that Baidu and Google simply determine relevant results differently than Bing and DuckDuckGo.

Lastly, we compared the six topics across the Bing/Baidu and Bing/DuckDuckGo comparisons of AnchorMAP and RBO results to see if there was any topics of agreement between these two pairings. The results of statistical tests showed that DuckDuckGo-Bing and Bing-Baidu have differing rankings across all topics and did not vary across topics.

8 FUTURE WORK

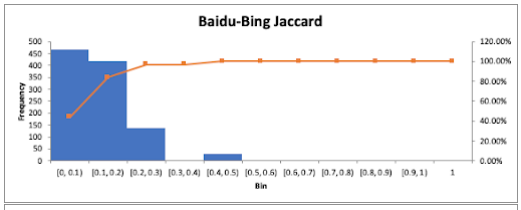
The consensus ranking procedure could be more robust than the methodology used in this paper. We employed a brute-force method to find each permutation of every combination of rankings, picking the ranking with the highest similarity to a pair of search engines. If there were two rankings that gave equivalent similarity values, then there were two possible consensus rankings, but our procedure only returned one ranking. It would be interesting to find these multiple consensus rankings and investigate the differences between them. The underlying consensus ranking procedure also could improved. The idea of a consensus or ground truth is an unsolved research question. We could explore the different approaches in the research community for determining consensus that are less naive than our approach.

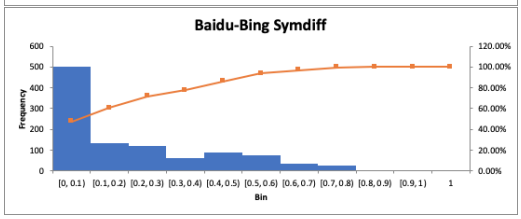
Our dataset could've been larger as well. Due to the constraint of time, we kept the datasets relatively small for the purposes of this study. However, the dataset still provides a quick snapshot across several pertinent topics and may be useful for future projects. The dataset, although somewhat dense, could not capture search engine behavior comprehensively. With more queries, we could possibly have a much denser and detailed distribution to analyze.

It would also be interesting to look into more search engines, especially ones that are used in other countries like Yandex or Naver. We could also perform the analysis on search engines from the same countries like Baidu, Sogou, and So.

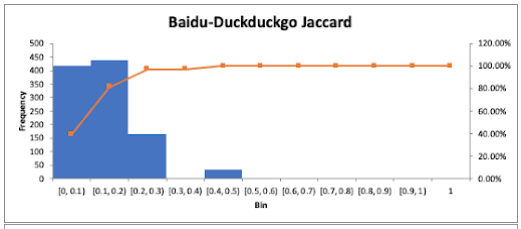
A APPENDIX

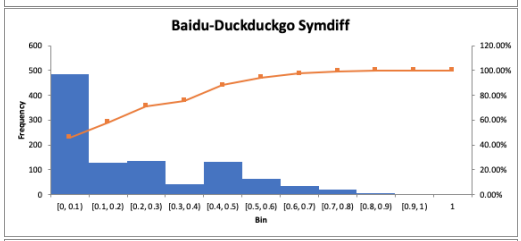
A.1 Comparisons between Baidu and Bing

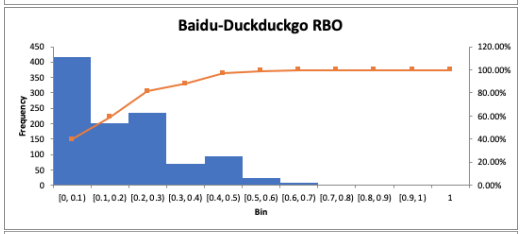


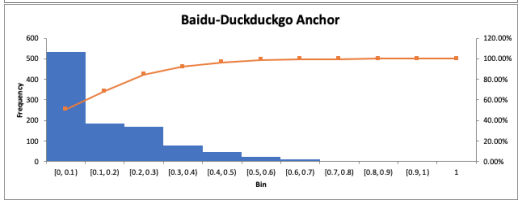


A.2 Comparisons between Baidu and DuckDuckGo

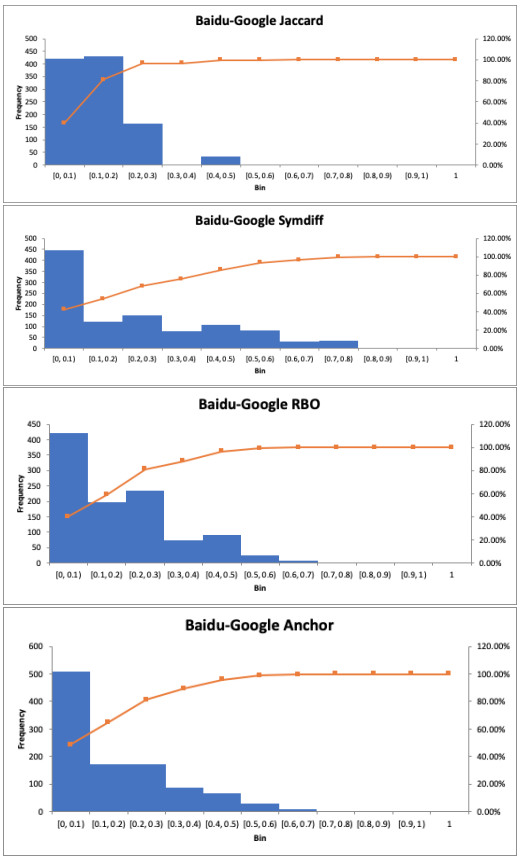




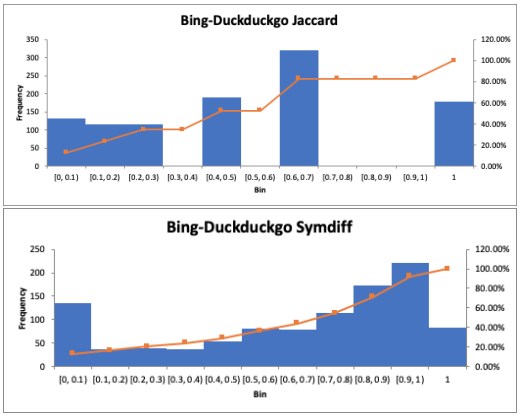




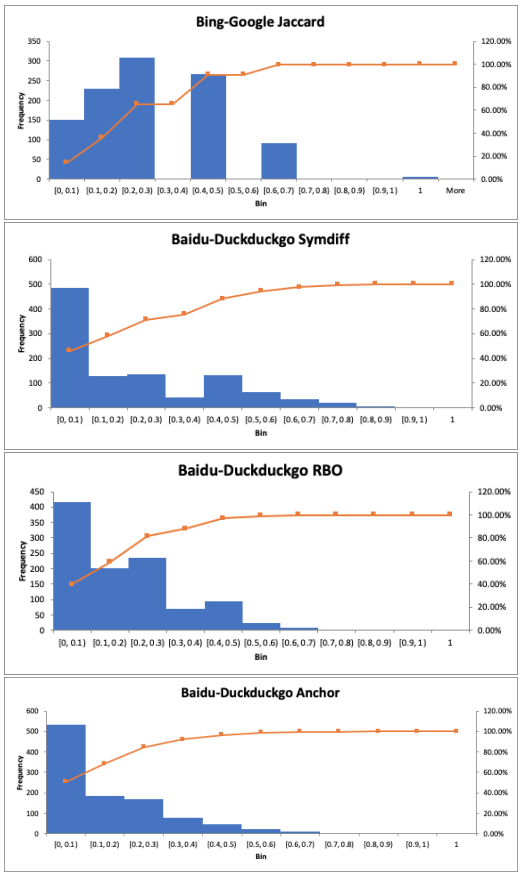
A.3 Comparisons between Baidu and Google



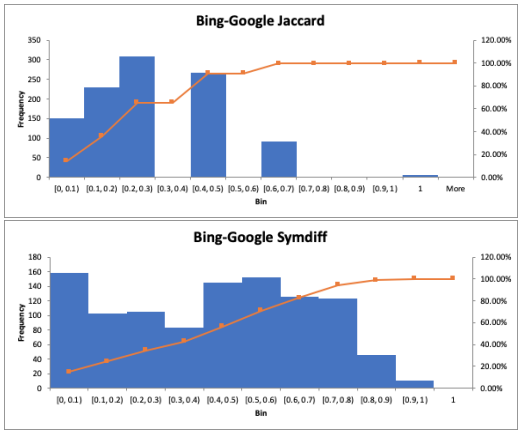
A.4 Comparisons between Bing and DuckDuckGo



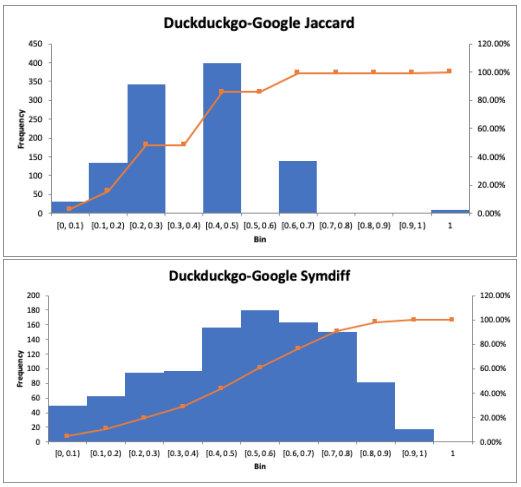
A.5 Comparisons between Baidu and DuckDuckGo



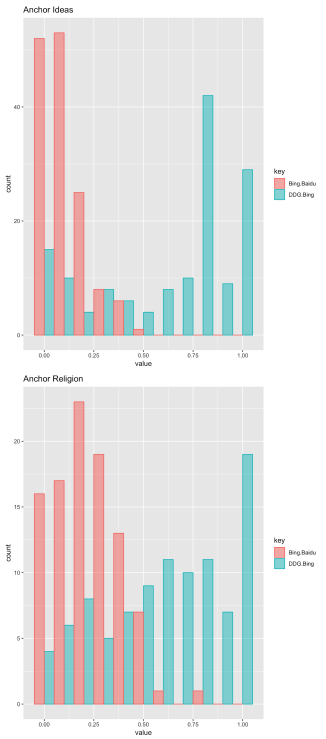
A.6 Comparisons between Bing and Google



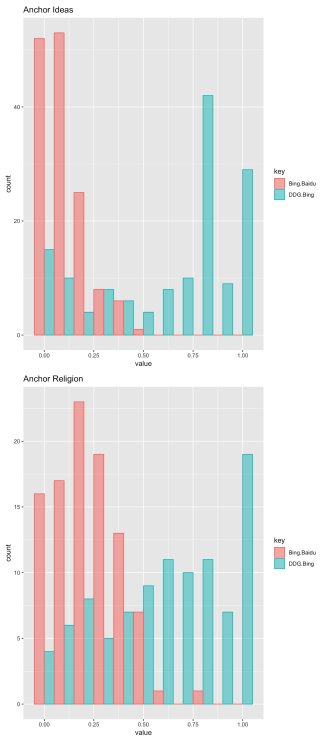
A.7 Comparisons between DuckDuckGo and Google

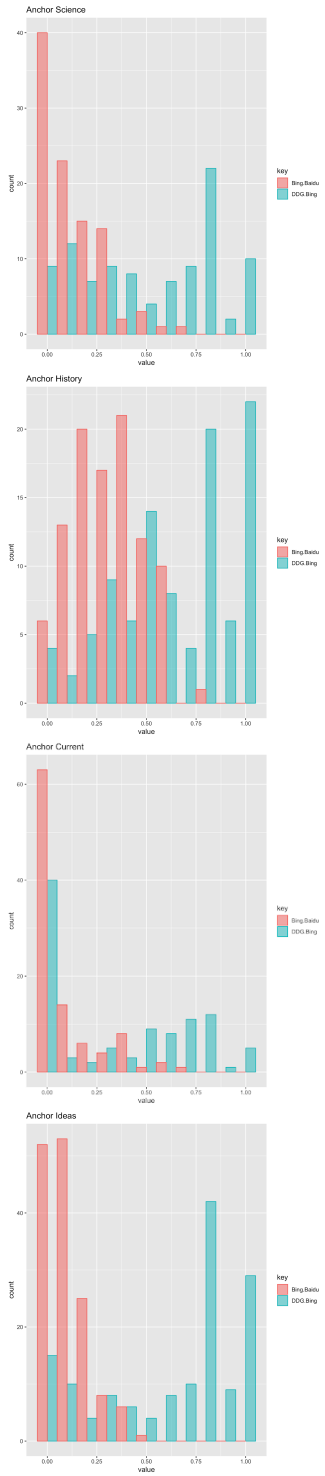


A.8 AnchorMAP across topics between Bing-Baidu and DuckDuckGo-Bing



A.9 RBO across topics between Bing-Baidu and DuckDuckGo-Bing





REFERENCES

[1] [n. d.]. Religious Tolerance Terms. http://www.religioustolerance.org/glossary.htm

[2] [n. d.]. Scraper API. https://www.scraperapi.com/

[3] Chris Buckley. 2004. Topic Prediction Based on Comparative Retrieval Rankings. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. ACM, New York, NY, USA, 506–507. https://doi.org/10.1145/1008992.1009093

[4] Błażej Ciepłuch, Ricky Jacob, Peter Mooney, and Adam C Winstanley. 2010. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*. University of Leicester, 337.

[5] Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing Top K Lists. (2003), 28–36. http://dl.acm.org/citation.cfm?id=644108.644113

[6] Zhiwei Lin, Yi Li, and Xiaolian Guo. 2017. Consensus of rankings. *CoRR* abs/1704.08464 (2017). arXiv:1704.08464 http://arxiv.org/abs/1704.08464

[7] Marina Meila, Kapil Phadnis, Arthur Patterson, and Jeff A. Bilmes. 2012. Consensus ranking under the exponential model. *CoRR* abs/1206.5265 (2012). arXiv:1206.5265 http://arxiv.org/abs/1206.5265

[8] NikolaiT. 2018. GoogleScraper. https://github.com/NikolaiT/GoogleScraper.

[9] Emily O'Neal. 2017. Quizlet. https://quizlet.com/203019548/us-history-flash-cards/.

[10] Rachael Tatman. 2018. Unigram Frequency. https://www.kaggle.com/rtatman/english-word-frequency.

[11] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (Nov. 2010), 38 pages. https://doi.org/10.1145/1852102.1852106

[12] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 587–594. https://doi.org/10.1145/1390334.1390435