Search Engine Similarity Analysis: A Combined Content and Rankings Approach

Konstantina Dritsa(✉), Thodoris Sotiropoulos(✉), Haris Skarpetis(✉), and Panos Louridas(✉)

Athens University of Economics and Business, Athens, Greece
{dritsakon,theosotr,p3110180,louridas}@aueb.gr

Abstract. How different are search engines? The search engine wars are a favorite topic of on-line analysts, as two of the biggest companies in the world, Google and Microsoft, battle for prevalence of the web search space. Differences in search engine popularity can be explained by their effectiveness or other factors, such as familiarity with the most popular first engine, peer imitation, or force of habit. In this work we present a thorough analysis of the affinity of the two major search engines, Google and Bing, along with DuckDuckGo, which goes to great lengths to emphasize its privacy-friendly credentials. To do so, we collected search results using a comprehensive set of 300 unique queries for two time periods in 2016 and 2019, and developed a new similarity metric that leverages both the content and the ranking of search responses. We evaluated the characteristics of the metric against other metrics and approaches that have been proposed in the literature, and used it to (1) investigate the similarities of search engine results, (2) the evolution of their affinity over time, (3) what aspects of the results influence similarity, and (4) how the metric differs over different kinds of search services. We found that Google stands apart, but Bing and DuckDuckGo are largely indistinguishable from each other.

Keywords: Search engines · Distance metrics · Results ranking · Document similarity

1 Introduction

Search engine battles make headlines in the international media; changes in their algorithms have become topics of business analysts. Their rollout is eagerly followed across the globe, while their inner workings remain corporate secrets.

The battle for prevalence in the search engine market is an ongoing game. Recent developments, such as the advent of stricter data protection policies, have affected the dynamics of the market. The United States search engine market

K. Dritsa and T. Sotiropoulos—These authors contributed equally to this work.

Z. Huang et al. (Eds.): WISE 2020, LNCS 12343, pp. 21–37, 2020.
https://doi.org/10.1007/978-3-030-62008-0_2

developments over the last three years show an increase of Google's market share by 5.45%, a decrease of Bing's market share by 18.13%, while DuckDuckGo's market share rose almost by a factor of four [18].

Beyond the comparative evolution of search engines, the similarity between search engines' results has been a topic of interest as it widely affects users' exposure to diverse or similar views and perspectives, especially for informative search [2]. There are two different approaches for comparing search engines' results: (1) ranking-based approaches that consider only the ordering of web results, and (2) content-based approaches that exploit only the textual content (i.e., snippets) of web responses. However, search engines are evolving at a fast pace, returning far richer results than the "ten blue links" of the past [30] and their evolution has given prominence to new user interaction patterns [10]. As a result, while the existing approaches can still be used for search engine comparisons, they are essentially a first-order approximation of the problem that does not take into account the current heterogeneous user experience.

In this work we tackle the question of the similarity between Google, Bing, and DuckDuckGo, by investigating whether and how their search results are different. For our comparison, we propose a novel similarity metric that takes into account both the top k lists [12,17] of search results, and their semantic content, as shown by the titles and text snippets in their responses. We apply our metric to a comprehensive set of queries gathered from two time periods.

Contributions. Our work contributes to both search engine affinity analysis and the top k results literature.

- **A novel metric for search engine similarity:** We introduce a combined content and rankings approach that returns more expressive similarity scores and distinguishes important differences in search engine behavior that are not apparent using the existing metrics.
- **Search engine affinity:** We develop an experimental setting for assessing the affinity of search engines. By assembling a varied set of 300 unique queries and inspecting their top 10 results over two distinct periods, one in 2016 and one in 2019, we compare the behavior of different search engines across time.
- **Comparison findings:** While Google appears to be different than both Bing and DuckDuckGo, the last two are indistinguishable from each other.

The rest of the paper is organized as follows: Section 2 provides an overview of related work. We introduce our metric in Sect. 3 and its application on our data set in Sect. 4. Section 5 presents our conclusions and further discussion.

2 Background and Related Work

The issues of affinity, performance, and stability in search engines have attracted research attention since their early days in the 1990s. The oldest studies [11, 14,19] focused mainly on evaluating and comparing the performance of search engines, employing a few queries (2 to 20) and manually examining the relevance of the results with the queries. In 2004, Google and two defunct search engines were evaluated, with Google demonstrating the best performance [28].

On the affinity of search results, studies until the late 2000s indicated a low overlap with mostly unique results [5,7,14]. In a 2010 study, Zaragoza et al. [33] conducted an alternative approach with quantitative statements, on 1000 queries in Google, Microsoft Live Search, and Yahoo! Search. The three search engines gave satisfactory results for navigational queries (i.e., queries that referred to a particular web page or service) and for frequent non-navigational queries.

At the same time, Webber et al. [31] developed Rank-Biased Overlap, a similarity metric for ranked lists. The researchers created a set of 113 queries and inspected the top 100 URLs produced by 11 search engines. Google and Microsoft Live Search results were common by 25%. Moreover, when checking against the localized versions of the search engines (e.g., the .au domain), Google was found to use less localization than Yahoo and Microsoft Live.

In a subsequent work in 2011 investigating the ranking similarity between Bing and Google [9], Cardoso and Magalhães applied the Rank-Biased Overlap on the results of 40,000 queries, showing that the search engines differed considerably. Furthermore, they looked into the diversity of search results for a given query using the Jensen-Shannon divergence and came to the conclusion that Bing tended to interpret a given query more diversely than Google.

In 2014, Collier and Konagurthu [12] proposed a measure for the comparison of two ranking lists, based on the minimum length encoding framework developed by Wallace [29]. The investigators measured the similarity between Ask, Google, and Yahoo for up to the top 100 results of 250 queries. Their findings showed that the search engines results differed linearly on their ranks, or quadratically using the Spearman and Kendall distances. Agrawal et al. [1] proposed two methods, TensorCompare and CrossLearnCompare, to compare search engine affinity, and used them to compare Google and Bing. We will return to this study in Sect. 4.6.

Although semantic features are largely incorporated into the process of producing and ranking search results, they are not integrated in the commonly used rank-distance metrics [8]. In cases where the results of the search engines have very similar URLs and rankings but the snippets/titles differ, a rank-distance metric cannot reflect this dissimilarity. On the other hand, approaches that solely focus on the content of web results would not sufficiently represent reality, when comparing search engines with similar snippets/titles, but different rankings.

In addition, the trend towards aggregation of multiple information sources into search results has led to changes in the corresponding evaluation methodologies [3,30]. Studies highlight interesting user interaction patterns [10] where the ordering of search responses does not play the sole role in browsing result pages. Research has shown that snippets and titles *notably* affect the user's decision to click on a specific page [13,24,26].

Unlike previous work, we propose a metric tailored to the search engine similarity problem that leverages diverse criteria as to the rankings and the content of web results. Our combined metric aims to return more expressive, objective, and robust similarity scores, highlighting differences that are not apparent from the existing metrics. Our metric also views each search result as it is; a unified piece of information. Furthermore, to the best of our knowledge, none of the prior studies include privacy-friendly search engines.

3 The Metric

We introduce a new metric, which we call T, to study search engine similarity. In Sect. 3.1 we formulate the problem that the metric aims to resolve and the criteria that it should meet; in Sects. 3.2–3.5 we develop metric T step-by-step. Then, in Sect. 3.6 we compare it to other existing metrics.

3.1 Problem Formulation

In what follows, we assume that for two search engines A and B we have two lists $R_A = [a_1, a_2, a_3, \ldots, a_n]$ and $R_B = [b_1, b_2, b_3, \ldots, b_n]$ of the ranked top n results of search engine A and search engine B respectively. We denote the i^{th} element of R_A with $R_A[i]$, and similarly for R_B.

Typically, search engine results consist of a URL, a result title, and a snippet describing the page content. Snippets and titles significantly affect the user's decision to click on a specific page [24,26]. To accurately appraise engine similarity, search engine comparisons should consider all these three aspects.

Motivating Example. To further highlight the importance of snippets and titles, consider Table 1 that shows the top result returned by Google and Bing for the query "Steven Wilson". Although search engines agree in the ordering of the same URL, they produce completely different snippets. The snippet produced by Bing focuses on artist's favorite film directors, while the snippet of Google gives emphasis on music news. Depending on user's search criteria, one snippet might be more effective on attracting user clicks than the other one.

Table 1. The top result retrieved for the query "Steven Wilson" on April 16, 2019.

	Bing	Google
Position	1	1
URL	http://stevenwilsonhq.com/sw/	http://stevenwilsonhq.com/sw/
Snippet	Steven is a film aficionado, and frequently cites cinema as one of the key inspirations for his music. Some of this favourite directors include Stanley Kubrick, David Lynch, Ben Wheatley, Jonathan Glazer, Shane Meadows and Christopher Nolan	The official website for songwriter/producer Steven Wilson. New live album/film 'Home Invasion: In Concert at the Royal Albert Hall' is out now!

Criteria. As the ranking of results does not fully capture their similarities, we need a comprehensive affinity metric that should meet the following criteria:

C1 The number of common elements (results). The more elements search engine A and B share in their top n results, the more similar they are.
C2 The distance of common elements. If an item appears in the results of both A and B, the affinity of A and B decreases as the distance of the element in the two result lists increases.
C3 The importance of agreement decreases as we go down in the results lists. For example, agreement at the top result is more important than that at the third or fourth result.
C4 If two search engines are similar, they produce similar titles and snippets, apart from returning similar results in a similar order.

3.2 Starting Point

As a starting point to define a metric for search engine affinity, we take the Jaro-Winkler distance, a variant of the Jaro distance [21], whose goal is to compute string similarity based on the common elements and the number of transpositions between them [32]. The Jaro distance of two strings S_1 and S_2 is given by:

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{1}$$

In the above, m is the number of matching characters and t denotes the number of transpositions. Two characters are considered matching if they are the same and their positions do not differ by more than $(\max(|S_1|, |S_2|)/2)-1$. The number of transpositions is defined as half the number of matching characters that are in different order in the two strings.

The Jaro-Winkler distance extends the Jaro distance by boosting it using a scaling factor p when the first l characters match exactly:

$$d_w = d_j + (l \times p \times (1 - d_j)) \tag{2}$$

In order to take into account the snippets and titles returned by the search engines, we adjust the Jaro-Winkler distance as follows:

$$S = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3n+1}\left(3m + 1 - a \cdot s - b \cdot h - c \cdot t\right) & \text{otherwise} \end{cases} \tag{3}$$

where n denotes the common length of the two result sets, $m = |R_A \cap R_B|$ is the number of common elements, t is the penalty from transpositions, s is the penalty from the differences between snippets, h is the penalty from the differences between titles, and a, b, $c \in [0, 1]$ are weights attached to the penalties accrued from snippets, titles, and transpositions respectively. To avoid division by zero, we add one to denominator.

Note that we compute the ratio of penalties $m - a \cdot s$ and $m - b \cdot h$ to the length n of results lists rather than the number of matching elements m, which is proposed by Jaro's metric. This gives us a more reliable estimation of the affinity between lists. For example, suppose we compare a pair of result rankings of length $n = 10$ and we get the number of matching elements as $m = 2$. According to Eq. 1, if $t = 0$ then the term $\frac{m-t}{m}$ is equal to 1 and it contributes $\frac{1}{3}$ to the overall similarity, which is a high number, considering the low number of matching items (two). Also, we use m/n instead of $m/|S_1| + m/|S_2|$, as R_A and R_B have a common length n.

3.3 Calculation of Penalties

Transpositions. To compute transpositions, we take the sum of the absolute differences of the positions of elements appearing in both lists. This is a variation of the deviation distance described by Ronald [25]. For lists R_A and R_B, the penalty is computed as follows, where $\sigma(R, e)$ is the position of e in list R:

$$t = \frac{\sum\limits_{e \in R_A \cap R_B} |\sigma(R_A, e) - \sigma(R_B, e)|}{t_{\max}}$$

This penalty is normalized on its upper bound. It can be proven that in the case of two lists of length n the upper bound for transpositions of $|R_A \cap R_B|$ is:

$$t_{\max} = \sum_{i=1}^{|R_A \cap R_B|} \phi(i, n)$$

where

$$\phi(i, n) = \begin{cases} n + 1 - i, & \text{if } i = 2k, k \in \mathbb{Z}^* \\ n - i, & \text{otherwise} \end{cases}$$

Snippets and Titles. The process of evaluating the penalties related to snippets and titles is common for both. We examine the sentences S_1, S_2 of snippets and titles that are produced by search engines A and B for a shared result. Then, we tokenize sentences S_1, S_2 and eliminate all stopwords as well as query terms. We get the union of all tokenized words that appeared in the two sentences and calculate the corresponding frequencies, forming two vectors V_1, V_2, that represent the actual snippets or titles. We then compute the cosine distance of the two vectors $d_s = 1 - \cos(V_1, V_2)$. The overall penalty is computed by iterating and repeating this process for all common results and summing all distances.

3.4 Similarity Boosting

The Jaro-Winkler metric treats all explicit matches at the first l characters of strings equally (recall Eq. 2). We, however, require a descending significance for agreement as we go down the list of results. To do that, we increase the value of S (Eq. 3) using weights w_i when there are common results in positions $1 \leq i \leq r \leq n$, with $w_1 > w_2 > \ldots > w_r$. This follows our third criterion, that exact or adjacent matches are more important at the beginning of results lists rather than the end. Moreover, in contrast to the Jaro-Winkler metric, the increase is not determined solely by the length of the matching prefix.

3.5 The Metric *T*

The final metric of similarity T combines the number of overlapping results as well as ordering, snippets, and titles of results, and it is given by:

$$T = S + \sum_{i=1}^{r} x_i w_i (1 - S) \tag{4}$$

Table 2. Results for the comparisons described in Sect. 3.6.

	`a b c d e f`
	`a g h i j k`
		`a b c d e f`
		`a b c g h i`
			`a b c d e f`
			`g h i d e f`
				`a b c d e f`
				`d e f a b c`
					`a b c d e f`
					`a b c d f e`
						`a b c d e f`
						`a b c f e d`

Spearman's footrule	1.0	1.0	1.0	0.0	0.89	0.78
Kendall's tau	1.0	1.0	1.0	0.85	0.98	0.95
G	0.29	0.71	0.57	0.29	0.95	0.90
M	0.48	0.82	0.36	0.18	0.98	0.97
Jaro-Winkler	0.44	0.77	0.67	0.0	0.96	0.96
T	$[0.24, 0.33]$	$[0.46, 0.68]$	$[0.25, 0.55]$	$[0.32, 0.48]$	$[0.59, 0.89]$	$[0.57, 0.88]$

where

$$x_i = \begin{cases} 0, & \text{if } R_A[i] \neq R_B[i] \\ 1, & \text{otherwise} \end{cases}$$

T meets all four criteria of Sect. 3.1. The calculation of overlapping items, m, fulfils C1. The computation of the penalty t fulfils C2, whereas boosting satisfies criterion C3. Finally, $a \cdot s$ and $b \cdot h$ cover C4.

3.6 Comparison with Other Metrics

In order to evaluate the behavior of our metric, we use a synthetic example and the criteria defined in Sect. 3.1 to contrast it with other metrics. Specifically, we compare it with Spearman's footrule and Kendall's tau (modified to measure similarity instead of distance) [17], the Jaro-Winkler metric, and the metrics G and M proposed by Bar-Ilan et al. [4,5].

Let $L_1 = [$`a b c d e f`$]$, a list that contains responses provided by one search engine. We compare L_1 with six other results lists $L_2 \ldots L_7$ using different metrics, as shown in Table 2. For the Jaro-Winkler metric we set $p = 0.1$, $l \leq 3$. In metric T we set $a = b = c = 1.0$ to penalize differences stemming from snippets, titles, and transpositions respectively, while we set $r = 3$, $w_1 = 0.15$, $w_2 = 0.1$, $w_3 = 0.05$ to reward matches at the first r elements.

Only metric T meets criterion C4 regarding snippets and titles. Thus, we present a lower and upper bound of our metric for every comparison. The lower bound corresponds to completely different snippets and titles among common results. The upper bound corresponds to snippets and titles that are identical.

In Table 2 we see that Spearman's footrule and Kendall's tau ignore mismatching elements, and compute similarity using only the common ones along with their distance, therefore, they do not meet criteria C1 and C3.

The Jaro-Winkler metric treats equally the transpositions of $d \leftrightarrow f$ and $e \leftrightarrow f$ in the comparisons (L_1, L_6) and (L_1, L_7) respectively, even though the former introduces a greater misplacement of elements. Thus, it violates criterion C2. Moreover, according to Eq. 2, it does not assign descending significance to agreements at the prefix of lists, which is required by criterion C3.

Both G and M metrics (the M metric to a greater extent) estimate the similarity of lists with emphasis to the ranking of items rather than the number of overlapping results. For example, we notice that even though L_1 and L_5 share all elements, the values of M and G show a decreasing importance to greater ranks, especially at the tail of lists. Also, a match in the first position, as in comparison (L_1, L_2), contributes 0.48 to the overall similarity according to the M metric, which is a great proportion relative to the number of matching items, i.e., *only* one out of total six. In essence, while M and G satisfy criteria C1–C3, they actually ignore matches or adjacent matches at the end of lists; in fact, the metric T can subsume G and M by using only the first q elements, for $q < n$.

Kumar and Vassilvitskii have proposed generalized versions of Spearman's footrule and Kendall's tau distances [22]; their versions take into account element weights, position weights, and element similarities in their calculations. It can be shown (omitted for reasons of space) that the generalizations overlook elements that appear only in a single list and thus miss criterion C1.

4 Evaluation

We compare Google, Bing, and DuckDuckGo (hereafter DDG), for numerous categories of queries, using our metric T. Google and Bing are the two dominant search engines, and have been the subject of comparative research. DDG adopts a different philosophy, placing a premium on user privacy. In our empirical evaluation, we try to answer the following research questions:[1]

RQ1 Do search engines produce similar web results? (Sect. 4.2)
RQ2 Is the similarity between search engines consistent over time? (Sect. 4.3)

[1] All data, results, and source code used on our experiments are available through https://doi.org/10.5281/zenodo.3980817.

RQ3 Which aspect of web results (i.e., rankings or content) influences the similarity of search engines the most? (Sect. 4.4)

RQ4 Do search engines produce similar results for different kinds of search services (i.e., news search service)? (Sect. 4.5)

RQ5 How do the results produced by the metric T correlate with the state-of-the-art? (Sect. 4.6)

4.1 Dataset

Table 3. Query categories

Books & Authors	Drinks & Food	Multinational companies	Music & Artists	Politicians
Regions	Software technologies	Sports	TV & Cinema	Universities

Our dataset consists of around 27600 top-10 result lists, spanning 10 categories of queries (Table 3). Each category contains around 30 queries; from these, 20 where taken from the U.S. version of Google Trends[2] in May 2016 and the rest were selected by us. Given that we cannot test all possible queries, we selected queries that affect a large number of users. For the data collection, we used the Bing Web Search API[3], the Google Custom Search API[4], and a web scraper that we developed for DDG. Our approach ensures that the search engines do not take user history into account, which would affect the final results [20]. We performed the queries daily, at the same time, using the American domain of each engine, for a period of one month (July–August) in 2016 and a period of 2 months (May–July) in 2019. We use both datasets to answer RQ2; for the rest of the research questions, both datasets gave consistent results, so, for brevity, we will focus on the 2019 dataset here.

Each result contains a URL, specifying its web location. Two identical URLs refer to the same result but a result could be pointed to by two different URLs [6]. To alleviate this issue, we applied standard normalization techniques [23] and resolved redirect HTTP responses to obtain the final target URL.

4.2 RQ1: Similarity of Search Engines

We estimate the similarity between Google, Bing, and DDG by employing metric T. For each search engine pair, we create a two-dimensional array D of the result similarity for every query and date. Each element D_{ij} represents the similarity between the two search engines in the day i for the query j.

[2] https://www.google.com/trends/topcharts.

[3] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/.

[4] https://developers.google.com/custom-search/.

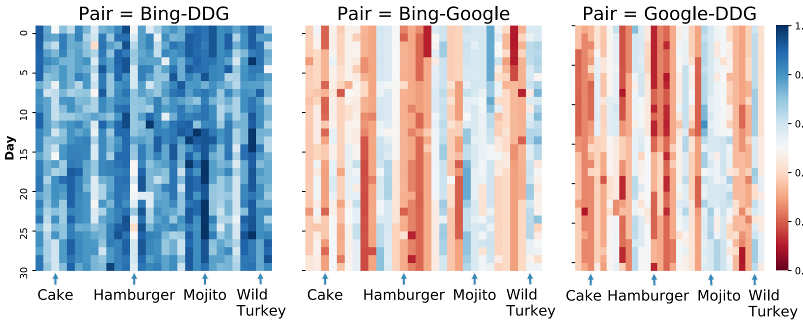

Fig. 1. Heatmaps of the DDG-Bing, Bing-Google and Google-DDG comparisons. The similarity between DDG and Bing is remarkable, while Google stands out. (Color figure online)

Recall from Eq. 4 that we need to define r and the weights $w_1, w_2, \ldots, w_r$ in order to reward matches at the first r elements of the ranking lists. In our experiments, we set $r = 5$ and $W = \{0.15, 0.1, 0.07, 0.03, 0.01\}$; we observed similar tendencies for different weight assignments. Regarding the importance of result factors, i.e., snippets, titles, and transpositions, we set $a = 0.8$, $b = 1$, $c = 0.8$. We use $b = 1$ as the weight for title penalties, because differences in titles are rare and in this way we could boost this factor (see Sect. 4.4).

Figure 1 presents the heatmaps of the similarity arrays D for the queries of the "Drinks & Food" category. These heatmaps are representative of all the other categories. Blue cells indicate cases where search engines are close to each other, while red cells reveal dissimilar web results.

We can see that Bing and DDG give very similar results for the vast majority of the queries. Despite its tiny market share, DDG still manages to offer a product comparable to that of the market leaders. The high Bing-DDG similarity could be explained by the fact that DDG -among other things- employs Bing to get its results [15]. Moving to Google-Bing and Google-DDG, the results of metric T indicate clear differences. However, there is still a number of queries where the search engines seem to have a high degree of resemblance, e.g., "Wild Turkey".

Finding #1. Google stands apart from Bing and DDG for the majority of the queries, while the latter two are mostly identical to each other.

4.3 RQ2: Consistency of Search Engines

To estimate the consistency of search engine behavior over time, we calculate the pair-wise average similarity score of each day, as computed by metric T. Figure 2 presents the average similarity of every search engine pair in time. This figure clearly shows that the affinity of the search engines is almost constant

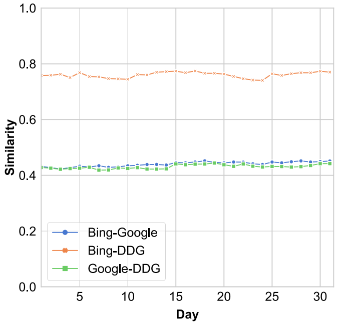

Fig. 2. Similarity evolution for a 31-day period in 2019. All pairs exhibit consistent behavior in the short-term.

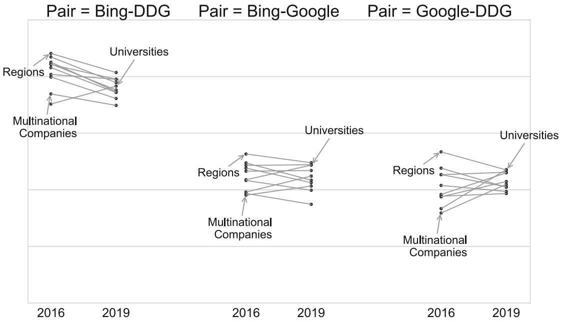

Fig. 3. Comparison of search engines' similarity in 2016 and 2019. The similarity does not change considerably in the long-term. The DDG-Bing similarity is almost double than the others.

over time, with only a small number of trivial fluctuations. The findings from this experiment imply that either the search engines do not significantly change their behavior or that their behavior changes in the same way. In addition, the plots reveal that the similarity between Bing-DDG is almost double than that of Bing-Google and Google-DDG, strengthening our first finding.

Finding #2. The behavior of all the search engines remains consistent in the short-term.

We also examined how the search engines' similarity changes in the long-term, by comparing the T similarities between the two time periods. Figure 3 shows the evolution of the pair-wise similarity for each query category from 2016 to 2019. Overall, we see that Bing and DDG moved from being very similar to slightly less so (their similarity decreases by 7.4%, on average). The affinity between of Bing-Google is almost stable (it drops by *only* 1.6%, on average), while DDG has come somewhat closer to Google, i.e., there is an increase in their similarity by 4.5%, on average. After inspecting the results, we found that these are due to changes in DDG's results within this time period.

Delving further into the data, we also examined how each search engine changed itself between these two points in time. We found that the average similarity between 2016 and 2019 is 0.37 for DDG, 0.43 for Bing, and 0.48 for Google; that is, Google's rankings and search algorithms changed the least and DDG has been updated to a greater extent, justifying its relative growth [18].

Finding #3. Bing and DDG remain more similar to each other than Bing-Google and Google-DDG. Although search engines change individually, their pairwise similarity is almost stable in the long-term.

4.4 RQ3: Impact of Snippets, Titles, and Transpositions

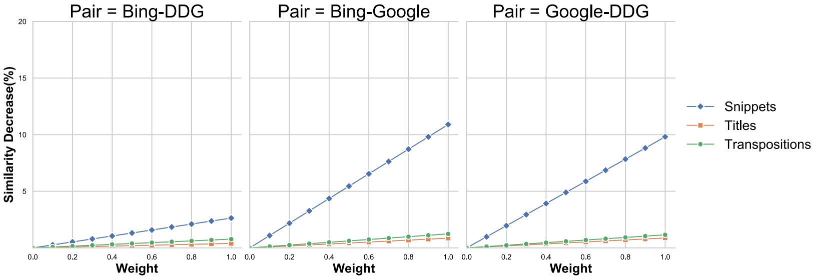

Fig. 4. The impact of snippets, titles and transpositions on each pair. The plots show the percentage decrease of similarity for various weight assignments of each factor. Google seems to construct different snippets compared to DDG and Bing.

Unlike existing approaches, metric T captures both the ordering (i.e., transpositions) and the content (i.e., snippets, titles) of results. Therefore, we can estimate how much each factor contributes to the differences of search engines. To do so, we instantiate the metric T with different weights for each factor (recall a, b, c from Eq. 3). We first consider the metric T_{base} as the baseline metric with weights $a = 0$, $b = 0$, $c = 0$. We compute the average similarity of every comparison pair for all the queries and days. Conceptually, T_{base} considers only the number of overlapping results and the agreements at the first $r = 5$ results. Then, we examine the effect of snippets by varying $a = 0.1, 0.2, \ldots, 1$ while keeping $b = c = 0$. Similarly, we examine the effect of titles and transpositions by varying b and c while keeping the other two weights pegged to zero.

In Fig. 4, each diagram shows the impact of every factor on the decrease of T_{base} for each search engine pair. It is clear that snippets have the biggest impact, while the difference in transpositions is much smaller, and in titles minimal. Google seems to construct different snippets compared to Bing and DDG, an observation that is consistent with our motivation example in Sect. 3.1.

Finding #4. Snippets have the greatest impact on the differences among all the comparison pairs; Google yields more distinct ones though. All the search engines tend to place their common results in adjacent positions. Finally, all the search engines produce almost identical titles.

4.5 RQ4: Search Engine Similarity in Different Search Services

Apart from standard web search, search engines provide a list of additional search services. We investigated whether our findings apply to the news search tab. We created a set of 30 news queries; 20 of them were taken from the Google News trends of May 2019 and the remaining 10 were generic news topics, e.g., “flood”.

The results show a low average similarity of 0.12, in contrast with the average 0.54 similarity of the results from the regular search. Furthermore, Bing-Google exhibits the highest similarity (0.15). This dissimilarity can be justified by the ephemeral nature of the news that requires quick evaluation, leading to daily ups and downs of topics and content. Also, the ranking algorithm of the news search results may be different than that of the regular search, certainly for Google [27].

Finding #5. There is a considerable difference in the results produced by different search engines’ services.

4.6 RQ5: Comparison with Other Approaches

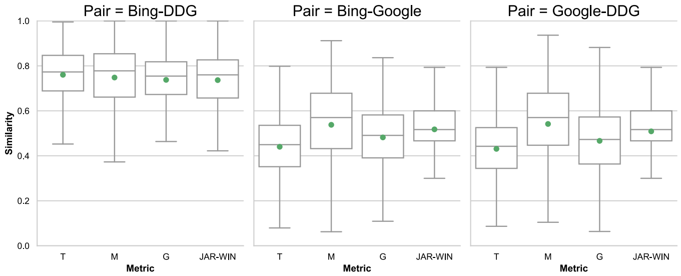

Fig. 5. Similarity of all search engine pairs using different metrics. Metric T exhibits lower box plots for the Bing-Google and Google-DDG comparisons, because it effectively captures the difference of their snippets and titles. (Color figure online)

Rankings-Based Approaches. We study how metric T correlates with three metrics that have been used in search engine comparisons (Sect. 3.6). Specifically, we use the metrics M, G, and Jaro-Winkler to compute the similarity between every search engine pair like we did in Sect. 4.2 using metric T.

Figure 5 shows the box plots of the search engine affinity for each metric. Every box plot contains the median similarity (horizontal line), the mean similarity (green circle), along with the maximum and minimum similarity values. The figure replicates our first finding (Sect. 4.2), that is, Google seems to produce more unique results when compared to Bing or DDG, as the corresponding box plots are lower than those of the Bing-DDG pair. Hence, the results of metric T are consistent with those of the three aforementioned metrics.

However, the box plots demonstrate that the metric T seems to distinguish from the others, especially in the Bing-Google, Google-DDG pairs. As shown by the average and the median similarity, metric T always produces lower values. This is explained by the fact that metric T *effectively* captures differences that stem from snippets and titles, which the other metrics ignore (Sect. 4.4).

Content-Based Approaches. Agrawal et al. [1] have proposed TensorCompare and CrossLearnCompare, two content-based methods that utilize tensor decomposition and supervised learning techniques. Both methods take into account the result snippets, but not their ordering. When applied on our data, the TensorCompare showed that the Bing-DDG pair is much more related than the rest, confirming our findings on their snippets similarity. The CrossLearnCompare, though, indicated an almost identical behavior for all search engines. An explanation for this could be that CrossLearnCompare actually predicts *queries* and not *search engines*, which may be distinguishable from each other. The Bing-DDG pair was more predictable than the rest, as we also find with metric T.

Finding #6. Metric T, when compared to others, exhibits a consistent behavior. However, when the content similarity falls, the results of metric T differ from those of the other metrics.

4.7 Threats to Validity

The main threat to external validity is the representativeness of the selected queries. To mitigate this threat, we created a large corpus of $27,600$ lists of top-10 search results, assembled from 300 unique queries, spanning 10 different topics. Two-thirds of the queries were taken from the Google trends of 2016, that impact a large number of users. The rest were selected by us, aiming to include less popular queries that better reflect the average search use. We considered only the top n web results for every query, as previous studies of user behavior [16] have shown it is more likely for users to click on one of the first ten results.

The main threat to internal validity is the design of our metric T. To alleviate this threat, we meet four criteria (Sect. 3.1) that are considered very important in search engine comparisons. When compared with existing approaches, our metric T demonstrates consistency with both ranking-based and content-based metrics. Another threat comes from the methodology of web results collection. We used the REST APIs of Google and Bing that do not consider user history [20]. We queried all search engines at the same time every day, with the same parameters and standard URL normalization methods.

5 Conclusions and Discussion

In this work, we introduce a novel similarity metric for search engine comparison that combines the rankings of results and their semantic presentation. In contrast to the existing ranking-based or content-based approaches, our metric aims to be more expressive, robust and objective, following the aggregation of heterogeneous information into search results and the emergence of new user interaction patterns. Thus, it effectively captures differences that stem from snippets and titles, which the other metrics ignore.

By employing our metric, we were able to track engine similarity on both content and ranking across time, for a large and broad number of queries. Our results indicate that Google stands apart from Bing and DuckDuckGo, but these two are largely indistinguishable. The performance of DuckDuckGo may run counter to many expectations, taking into account the comparatively vast disparity of its resources. In our study we queried search engines without taking into account the user history. It is possible that when user history is employed, Bing would differ measurably from DuckDuckGo. Still, Google manages to differ from both Bing and DuckDuckGo even when it does not leverage personalized data.

Lately, search engines have started producing summaries, overviews, and compelling navigational aids, calling for more flexible comparison methodologies. Our approach consists a first step towards this direction, but the incorporation of semantically-rich features in search engine similarity measures seems a promising area for future research.

Acknowledgments. This work was supported by the European Union's Horizon 2020 research and innovation program "FASTEN" under grant agreement No. 825328.

References

1. Agrawal, R., Golshan, B., Papalexakis, E.: A study of distinctiveness in web results of two search engines. In: Proceedings of the 24th International Conference on World Wide Web (2015)
2. Agrawal, R., Golshan, B., Papalexakis, E.: Whither social networks for web search? In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015)
3. Bailey, P., Craswell, N., White, R.W., Chen, L., Satyanarayana, A., Tahaghoghi, S.: Evaluating whole-page relevance. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2010)
4. Bar-Ilan, J., Levene, M., Mat-Hassan, M.: Dynamics of search engine rankings–a case study. In: WebDyn@ WWW (2004)
5. Bar-Ilan, J., Mat-Hassan, M., Levene, M.: Methods for comparing rankings of search engine results. Comput. Netw. **50**(10), 1448–1463 (2006)
6. Bar-Yossef, Z., Keidar, I., Schonfeld, U.: Do not crawl in the DUST: different URLs with similar text. ACM Trans. Web **3**(1), 1–31 (2009)
7. Bharat, K., Broder, A.: A technique for measuring the relative size and overlap of public web search engines. Comput. Netw. ISDN Syst. **30**(1), 379–388 (1998)
8. Bian, J., Liu, T.Y., Qin, T., Zha, H.: Ranking with query-dependent loss for web search. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (2010)
9. Cardoso, B., Magalhães, J.: Google, Bing and a new perspective on ranking similarity. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (2011)

10. Chen, D., Chen, W., Wang, H., Chen, Z., Yang, Q.: Beyond ten blue links: enabling user click modeling in federated web search. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (2012)
11. Chu, H., Rosenthal, M.: Search engines for the world wide web: a comparative study and evaluation methodology. In: Proceedings of the ASIS Annual Meeting, vol. 33 (1996)
12. Collier, J.H., Konagurthu, A.S.: An information measure for comparing top k lists. In: 2014 IEEE 10th International Conference on e-Science, vol. 1 (2014)
13. Cutrell, E., Guan, Z.: What are you looking for?: An eye-tracking study of information usage in web search. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2007)
14. Ding, W., Marchionini, G.: A comparative study of web search service performance. Proc. ASIS Ann. Meet. **33**, 136–142 (1996)
15. DuckDuckGo: DuckDuckGo sources (2019). https://help.duckduckgo.com/results/sources/. Accessed 07 Aug 2019
16. Enge, E., Spencer, S., Fishkin, R., Stricchiola, J.: The Art of SEO. O'Reilly Media, Inc., Sebastopol (2012)
17. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM J. Discrete Math. **17**(1), 134–160 (2003)
18. StatCounter GlobalStats: Statcounter globalstats (2019). http://gs.statcounter.com. Accessed 06 Aug 2019
19. Gordon, M., Pathak, P.: Finding information on the world wide web: the retrieval effectiveness of search engines. Inf. Process. Manag. **35**(2), 141–180 (1999)
20. Hannak, A., et al.: Measuring personalization of web search. In: Proceedings of the 22nd International Conference on World Wide Web. ACM (2013)
21. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J. Am. Stat. Assoc. **84**(406), 414–420 (1989)
22. Kumar, R., Vassilvitskii, S.: Generalized distances between rankings. In: Proceedings of the 19th International Conference on World Wide Web. ACM (2010)
23. Lee, S.H., Kim, S.J., Hong, S.H.: On URL normalization. In: Gervasi, O., et al. (eds.) ICCSA 2005. LNCS, vol. 3481, pp. 1076–1085. Springer, Heidelberg (2005). https://doi.org/10.1007/11424826_115
24. Maxwell, D., Azzopardi, L., Moshfeghi, Y.: A study of snippet length and informativeness: behaviour, performance and user experience. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (2017)
25. Ronald, S.: More distance functions for order-based encodings. In: 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98TH8360), May 1998
26. Sachse, J.: The influence of snippet length on user behavior in mobile web search. Aslib J. Inf. Manag. **71**(3), 325–343 (2019)
27. The Economist: Seek and you shall find: Google rewards reputable reporting, not left-wing politics, June 2019. https://www.economist.com/graphic-detail/2019/06/08/google-rewards-reputable-reporting-not-left-wing-politics
28. Vaughan, L.: New measurements for search engine evaluation proposed and tested. Inf. Process. Manag. **40**(4), 677–691 (2004)
29. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. Springer, New York (2005). https://doi.org/10.1007/0-387-27656-4
30. Wang, Y., et al.: Optimizing whole-page presentation for web search. ACM Trans. Web **12**(3), 1–25 (2018)

31. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. (TOIS) **28**(4), 1–38 (2010)
32. Winkler, W.E.: String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods (1990)
33. Zaragoza, H., Cambazoglu, B.B., Baeza-Yates, R.: Web search solved?: All result rankings the same? In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (2010)