The Journal of Web Science, 2016, 2: 17-30

Overlap in the Web Search Results of Google and Bing

Rakesh Agrawal[1], Behzad Golshan[2] and Evangelos Papalexakis[3]

[1] *Data Insights Laboratories, ragrawal@acm.org*

[2] *Boston University, behzad@bu.edu*

[3] *Carnegie Mellon University, epapalex@cs.cmu.edu*

ABSTRACT

Google and Bing have emerged as the diarchy that arbitrates what documents are seen by Web searchers, particularly those desiring English language documents. We seek to study how distinctive are the top results presented to the users by the two search engines. A recent eye-tracking has shown that the web searchers decide whether to look at a document primarily based on the snippet and secondarily on the title of the document on the web search result page, and rarely based on the URL of the document. Given that the snippet and title generated by different search engines for the same document are often syntactically different, we first develop tools appropriate for conducting this study. Our empirical evaluation using these tools shows a surprising agreement in the semantics of the results produced by the two engines for a wide variety of queries used in our study. Thus, this study raises the open question whether it is feasible to design a search engine that would produce results distinct from those produced by Google and Bing that the users will find helpful.

Keywords: Web search; search engine; search result comparison; web mining; Google; Bing

ISSN 2332-4031; DOI 10.1561/106.00000005

1 Introduction

The World Wide Web is now widely recognized as the universal information source. The fairness doctrine enunciated several decades ago contends that citizens should have access to diverse perspectives Federal Communications Commission, 1949. The normative impetus behind this doctrine is the idea that exposure to different views is beneficial for citizens. Without question, content representing diverse perspectives exist on the Web almost on any topic However, this does not automatically guarantee that audiences encounter them Stroud and Muddiman, 2012.

Search engines have become the dominant tool used to access the web content Purcell *et al.*, 2012. In the physical world, one way people gain access to diverse perspectives is by subscribing to different newspapers, listening to different radio stations, tuning into different television channels, or manually selecting different publications and books. We seek to study whether users can garner different perspectives by obtaining results for the same query from different search engines. For this purpose, we study how distinctive are the web search results produced by Google and Bing - the two most popular search engines of English language documents (Yahoo's web search is currently powered by Bing).

In addition to the information about the documents that the search engine deems most relevant to the query (the so called "organic results"), a search engine result page (SERP) often contains a variety of other information. This may include inter alia sponsored listings, images, videos, maps, definitions, or suggested search refinements. We focus on comparing the top-10 organic results on the first SERP because they are the ones that get almost all of the clicks Enge *et al.*, 2012. Users unsatisfied with the top results frequently retype a query, instead of looking at results at lower positions Guan and Cutrell, 2007. An organic result normally includes the title of the document, a snippet of the document, and URL of the full version.

A recent eye-tracking has shown that the web searchers decide whether to look at a document primarily based on the snippet and secondarily on the title of the document on the web search result page, and rarely based on the URL of the document Marcos and González-Caro, 2010. Given that the snippet and title generated by different search engines for the same document are often different, we first develop tools appropriate for conducting this study. We then use these tools to study the extent of agreement in the results produced by the two engines for a wide variety of queries.

Contributions

In this work, our main contribution is quantifying how distinctive are the organic search results produced by Google and Bing. In order to achieve that, we also make the following technical contributions:[1]

[1] While designed to effectively analyze search engine results, our tools have broader applicability. For instance, consider a set of questions, possibly coming from a Massive Open Online Course (MOOC) exam; these questions can either be multiple choice or in free-text form. In the setting of a MOOC, there will be potentially hundreds, or thousands, of students responding to those questions. There are also multiple exams, as well as multiple MOOCs on the same subject, offered by different providers. Using these tools, we are able to quantify the similarity of students across different MOOCs, as well as similarity of MOOCs in terms of how students respond to exam questions (which could be an indicator of how well students learn from a particular MOOC).

Visualization and exploratory analysis: We introduce TensorCompare, an exploratory tool for visualizing and analyzing pairwise differences between search engines.

Quantitative comparison of search engines results: We also introduce CrossLearnCompare, a tool that uses machine learning and quantifies the similarity of results between two search engines by framing it as a prediction problem.

Paper Layout

The structure of the rest of the paper is as follows. We begin by discussing related work in Section 2. We then describe the new tools we designed for carrying out the comparative study in Section 3. Section 4 presents the empirical evaluation. We conclude with a discussion of the significance of the work and future directions in Section 5.

2 Related Work

More than four decades ago, Lancaster and Fayen Lancaster and Fayen, 1973 in 1973 listed six criteria for assessing the performance of information retrieval systems: 1) Coverage, 2) Recall, 3) Precision, 4) Response time, 5) User effort, and 6) Form of output. Since the advent of search engines in early 90's, there are several reported studies that evaluated their performance on one or more these criteria. See Chu and Rosenthal, 1996 and references therein for examples of some early studies. See Lewandowski, 2012 for a recent compilation of various issues and studies related to the evaluation of web search engines. We will focus our discussion on prior works that studied the overlap of results between different search engines, the thrust of our paper.

An early overlap study is due to Ding and Marchionini, who measured the result overlap between the then popular three search engines: InfoSeek, Lycos, and OpenText. Five queries were used to conduct searches with these services. They observed a low level of result overlap among the services Ding and Marchionini, 1996. Around the same time, Selberg and Etzioni found, in the context of their metacrawler work, that none of Galaxy, Infoseek, Lycos, OpenText, Webcrawler and Yahoo was able to return more then 45% of the references followed by users. They also observed that each of the engines returned mostly unique results Selberg and Etzioni, 1995. Also in 1996, Gauch, Wang and Gomez found that a metasearch engine that fused the results of Alta Vista, Excite, InfoSeek, Lycos, Open Text, and WebCrawler provided the highest number of relevant results Gauch and Wang, 1996.

Bharat and Broder estimated the size of the Web to be 200 million pages in November 1997 and the overlap between the websites indexed by HotBot, Alta Vista, Excite and InfoSeek to be only 1.4% Bharat and Broder, 1998. Lawrence and Giles published their study of AltaVista, Excite, HotBot, Infoseek, Lycos, and Northern Light in 1998. They found that the individual engines covered from 3 to 34% of the indexable Web, based on their estimate of the size of the Web at 320 million pages. Combining the six engines in their study covered about 3.5 times as much of the Web as one engine Lawrence and Giles, 1998.

Fast forwarding a bit, Gulli and Signorini estimated that by January 2005 the indexable Web had increased in size to about 11.5 billion pages and that Google's coverage rate was 76.2%, Yahoo's 69.3% and that of MSN Search (predecessor of Bing) 61.9% Gulli and Signorini, 2005. Spink et al. studied the overlap between the results of four search engines, namely MSN, Google, Yahoo and Ask Jeeves, using data from July 2005. Their findings showed that the percent of total first page results unique to only one of the engines was 84.9%, shared by two of the three was 11.4%, shared by three was 2.6%, and shared by all four was 1.1% Spink *et al.*, 2006. In an update two years later, they noted that the first page results of the four engines continued to differ from one another and in fact they included fewer results in common in 2007 than in 2005 Spink *et al.*, 2008.

More recently, Pirkola investigated how effectively the websites of Finnish, French, and U.S. domains were being indexed by two US-based and three Europe-based search engines Pirkola, 2009. The results showed that Google and Live Search (predecessor of Bing) indexed US sites more effectively than Finnish and French sites, the Finnish www.fi indexed only Finnish sites and the French Voila only French sites, and the European engine Virgilio indexed European sites more effectively than US sites. In another interesting study, Wilkinson and Thelwall compared the results of seventeen random queries submitted to Bing for thirteen different English geographic search markets at monthly intervals Wilkinson and Thelwall, 2013. They found there were almost no ubiquitous authoritative results: only one URL was always returned in the top-10 for all search markets and points in time and that results from at least three markets needed to be combined to give comprehensive results. There also have been studies pointing out that the search engine results are not stable even in short windows of time Bar-Ilan, 2004; Lewandowski, 2012.

We did not find much discussion in prior work of the techniques used for determining if two result pages contained links to the same web document. For example, Spink *et al.*, 2006; Spink *et al.*, 2008 simply state that this determination is done using string comparison of URLs. It is not clear what URL normalization Lee *et al.*, 2005; Lei *et al.*, 2010, if any, was done before string comparison. It is also not clear what, if anything, was done to address the problem of DUST - Different URLs with Similar Text Bar-Yossef *et al.*, 2009. Finally, there is no mention of short URLs, although the first notable URL shortening service, namely tinyURL, dates back to 2002 Antoniades *et al.*, 2011.

In our work, we use its snippet to represent a search result. Apart from the present work, we are aware of another work that uses snippets as a means of representing and comparing search results, albeit not focused on comparing Google and Bing. Specifically, Teevan, Ramage, and Morris Teevan *et al.*, 2011 (TRM Study) extracted snippets of the search results from the Bing search logs for 42 most popular queries for one week in 2009, and also obtained all the tweets containing those queries during the same period. They then computed per query average cosine similarity of each web snippet with the centroid of the other web snippets and with the centroid of the tweets. Similarly, they computed the per-query average cosine similarity of each Twitter result with the centroid of the other tweets and with the centroid of the web snippets. All averaging and comparisons are done in the reduced topic space

obtained using Latent Dirichlet Allocation (LDA) Blei *et al.*, 2003. We adopt this technique from the TRM study as an additional measure of similarity in order to increase confidence in our findings.

To summarize, all of prior work found little overlap between the first page results produced by different engines for very many queries. Some plausible reasons have also been put forward for this low overlap. They include that the search engines are constrained in the portions of the Web they index due to network bandwidth, disk storage, computational power, or a combination of these items. Search engines use different technologies to find pages and indexing them. And they deploy proprietary algorithms to determine the ranking of the results and their presentation to the users. Fingers have also been pointed at implicit personalization Hannak *et al.*, 2013.

Why another study?

Given the rich prior literature we have outlined, it is natural to question the need for a new study on the overlap between the search engine results. We believe that much has changed in the recent times in the search engine market landscape and in search engine technologies to warrant such a study. With Bing powering Yahoo search, we now essentially have a diarchy in Google and Bing that arbitrates user access to the English language Web that a very large fraction of humanity accesses on daily basis to get information. But there is no recent comparative study of Google and Bing search results. It is imperative to periodically analyze what people are able to see and read. Such studies also lead to the creation of new analysis tools and the questioning of conventional wisdom, thus contributing to the advancement of science.

3 Analytical Tools

We designed two tools to be able to analyze and compare search engine results. One, which we call TensorCompare, uses tensor analysis to derive low-dimensional compact representation of search results and study their behavior over time. The other, which we call CrossLearnCompare, uses cross-engine learning to quantify their similarity. Throughout the text, we use the term *semantic* in the same way as the highly influential Latent Semantic Indexing work Deerwester *et al.*, 1990 does, i.e., *"terms in a document may be taken as referents to the document itself or to its topic"*. We discuss the two proposed methods next.

3.1 TensorCompare

Postulate that we have the search results of executing a fixed set of queries at certain fixed time intervals on the same set of search engines. Suppose, further, that we have a vector representation of each particular set of organic results, in a feature space. For instance, that feature space could encode the set of URLs in the results, or a bag-of-words representation of the keywords appearing in the snippets of those results. We thus have a four-way relation of *queries*, *result features*, the *time* when these results were obtained, and the *search engine* that yielded those results. This four-way relation can be represented in a four mode[2] tensor $\underline{\mathbf{X}}$, where (query, result, time, search engine) are the four modes. A tensor is a higher order generalization of a matrix (and in the case of binary relations, a tensor is simply a matrix). We refer the interested reader to Kolda and Bader, 2009 for a thorough overview. In our case, $\underline{\mathbf{X}}$ can be binary valued or real valued (indicating, for instance, frequencies).

This tensor can be analyzed using the so-called canonical or PARAFAC decomposition Harshman, 1970, which decomposes the tensor into a sum of rank-one tensors: $\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \lambda_r \ \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \circ \mathbf{d}_r$, where the (i,j,k,l)-th element of $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} \circ \mathbf{d}$ is simply $\mathbf{a}(i)\mathbf{b}(j)\mathbf{c}(k)\mathbf{d}(l)$. The vectors $\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r, \mathbf{d}_r$ are usually normalized, with their scaling absorbed in λ_r. For compactness, the decomposition is represented as matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$. Computing the decomposition is a an intensive task. For an $I \times J \times K \times L$ tensor and for R components, the Alternating Least Squares algorithm (which is considered the work-horse algorithm for the PARAFAC decomposition) has complexity $O(IJKLR)$. However, there is significant work in exploiting sparsity Bader and Kolda, 2007a, algorithms designed for Map/Reduce Kang *et al.*, 2012, and more scalable approaches Papalexakis *et al.*, 2013; Sidiropoulos *et al.*, 2014, which enable decomposition for very large tensors. In Section 4.4, we include run-time measurements for our results.

The decomposition of $\underline{\mathbf{X}}$ to $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ gives us a low rank embedding of queries, results, timings, and search engines respectively. Each rank-one component $\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r, \mathbf{d}_r$ can be seen as a co-cluster that associates queries, results, timestamps, and search engines The largest values within each vector serve as the membership indicators for this co-cluster: for instance, the largest values of $\mathbf{a}_r$ will indicate which queries are contained in the r-th co-cluster. In the exemplar case where the results are represented in a bag-of-words feature space, each co-cluster will have a subset of the queries, a set of terms that are prominent for those queries and are semantically similar, a set of dates for which those result terms where produced, and finally a set of search engines that yielded those results. An alternative, equivalent view of each rank-one component is the one of an augmented topic model that contains information about the queries, the search engines, as well as the dates for which this topic was present.

Our primary goal is to semantically compare search engines, and thus we turn our attention to factor matrix $\mathbf{D}$. This matrix projects each one of the search engines to the R-dimensional space. Alternatively, one can view this embedding as soft clustering of the search engines, with matrix $\mathbf{D}$ being the cluster indicator matrix: the (i,j) entry of $\mathbf{D}$ shows the participation of search engine i in cluster j.

This leads to a powerful visualization tool that captures similarities and differences between the search engines in an intuitive way. Say we take search engines A and B and the corresponding rows of matrix $\mathbf{D}$. If we plot these two row vectors against each other, the resulting plot will contain as many points as clusters (R in our particular notation). The positions of these points are the key to understanding the similarity be-

[2] In the literature, "mode" refers to the aspects/modalities of the data (e.g., queries, or search engines) whose relations are represented by the tensor. We avoid using the term "dimension" because it usually refers to the size of each mode (e.g., the number of queries, or the number of search engines etc).

tween search engines.

Figure 1 serves as a guide. The (x, y) coordinate of a point on the plot corresponds to the degree of participation of search engines A and B respectively in that cluster. If all points lie on the 45 degree line, this means that both A and B participate equally in all clusters. In other words, they tend to cluster in the exact same way for semantically similar results and for specific periods of time. Therefore, Fig. 1(a) paints the picture of two search engines that are very (if not perfectly) similar with respect to their responses. In the case where we have only two search engines, perfect alignment of their results in a cluster would be the point $(0.5, 0.5)$. If we are comparing more than two search engines, then we may have points on the lower parts of the diagonal. In the figure, we show multiple points along the diagonal for the sake of generality.

Figure 1(b), on the other hand, shows the opposite behavior. Whenever a point lies on either axis, this means that only one of the search engines participate in that cluster. If we see a plot similar to this figure, we can infer that A and B are very dissimilar with respect to their responses. In the case of two search engines, the only valid points on either axis are $(0, 1)$ and $(1, 0)$, indicating an exclusive set of results. However, for generality, we show multiple points on each axis.

Note, of course, the cases shown in Fig. 1 are the two extremes, and we expect to observe behaviors bounded by those extremes. For instance, in the case of two search engines, all points should lie on the line $\mathbf{D}(1, j)x + \mathbf{D}(2, j)y = 1$, where $\mathbf{D}(1, j)$ is the membership of engine A in cluster j, and $\mathbf{D}(2, j)$ is the membership of engine B in cluster j. This line is the dashed line of Fig. 1(a).

Choosing the number of clusters R is a very interesting, open problem. Typically, R is chosen to be smaller than the rank of the tensor. However, determining that rank, unlike in the matrix case (where the Singular Value Decomposition reveals the rank of the matrix) is an NP-complete problem Håstad, 1990. Fortunately, there exist useful heuristics that can determine whether a given R is appropriate for the tensor at hand, such as the Core Consistency Diagnostic Bro and Kiers, 2003, and the Automatic Relevance Determination Mørup and Hansen, 2009.

TensorCompare also allows us to track the behavior of clusters over time. In particular, given the i-th group of semantically similar (query, result, search engine) cluster, as given by the decomposition, the i-th column of matrix $\mathbf{C}$ holds the temporal profile of that cluster. Suppose we have T days worth of measurements. If the search engines of that cluster produce similar results for the given set of queries for all T, the temporal profile will be approximately constant and each value will be approximately equal to $\frac{1}{T}$. Otherwise, there will be variation in the profile, correlated with the variation of the particular results. In the extreme case where a result appeared only on a single day, the time profile will have the value approximately equal to one corresponding to that day, and approximately zero for the rest of the days.

Theoretical Foundation

We next provide a Lemma that connects the plots provided by TensorCompare to the degree of semantic overlap of two search engines. Suppose that for a given cluster j, we denote the membership of search engine A as $x = D(A, j)$ and the membership of search engine B as $y = D(B, j)$. For ease of exposition, consider the case of two search engines and assume that we have a three mode tensor: (query, result, search engine).

Algorithm 1: CrossLearnCompare

Input: $\mathcal{R}_A$, $\mathcal{R}_B$ are instances of results of engines A and B. Each instance is in the form (query, result representation in chosen feature space)

Output: Similarity measures $c_{A,B}$ and $c_{B,A}$ between search engines A, B.

1: Train a model $\mathcal{M}_A$ based on the instances $\mathcal{R}_A$, using the query as a class label.
2: Train a model $\mathcal{M}_B$ based on the instances $\mathcal{R}_B$, using the query as a class label.
3: For all instances in $\mathcal{R}_B$, use $\mathbf{M}_A$ to predict the query. Set $c_{A,B}$ as a measure of the classifier's accuracy (e.g. Area Under the Curve).
4: For all instances in $\mathcal{R}_A$, use $\mathbf{M}_B$ to predict the query. Set $c_{B,A}$ likewise.

Lemma 1. *Assume a binary (query, result, search engine) tensor that has exactly one rank one component. Let search engine A correspond to the x coordinate, and search engine B correspond to the y coordinate of a* TensorCompare *plot. For the particular component, if search engine B has p_1 fraction of queries in common with A, and p_2 portion of the result in common with A, then*

$$y \leq p_1 p_2 x.$$

Proof. See Appendix. □

In the case of a four-mode tensor, with p_3 percent overlap in the time mode, the bound is $y \leq p_1 p_2 p_3 x$. The above Lemma provides an upper bound, however, we experimentally validated that this bound is in practice tight.

3.2 *CrossLearnCompare*

An intuitive measure of the similarity of the results of two search engines is the predictability of the results of a search engine given the results of the other. Say we view each query as a class label. We can then go ahead and learn a classifier that maps the search result of search engine A to its class label, i.e. the query that produced the result. Imagine now that we have results that were produced by search engine B. If A and B return completely different results, then we would expect that classifying correctly a result of B using the classifier learned using A's results would be difficult, and our classifier would probably err. On the other hand, if A and B returned almost identical results, classifying correctly the search results of B would be easy. In cases in between, where A and B bear some level of similarity, we would expect our classifier to perform in a way that it is correlated with the degree of similarity between A and B.

Note we can have different accuracy when predicting search engine A using a model trained on B, and vice versa. This, for instance, can be the case when the results of A are a superset of the results of B. Algorithm 1 shows an outline of CrossLearnCompare.

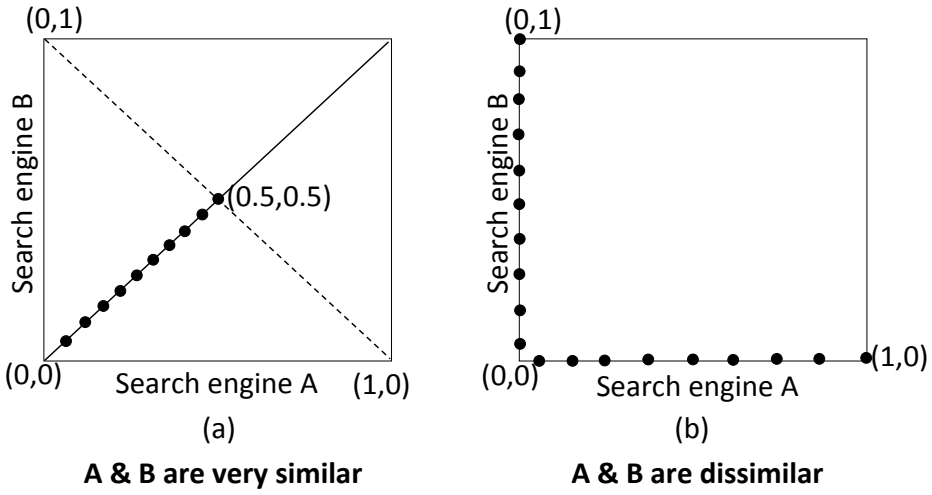


Figure 1: Visualization guide for TensorCompare.

4 Empirical Evaluation

We now present the results of the empirical study we performed, applying the tools just described on the search results from Google and Bing for a wide variety of queries.

4.1 Data Set

We conducted the evaluation for two sets of queries. The Trends set (Table 1) contains the most popular search terms from different categories from Google Trends during April 2014. We will refer to them as *head queries*. The Manual set (Table 2) consists of hand-picked queries by the authors that we will refer to as *trunk queries*. These queries consist of topics that the authors were familiar with and were following at the time. Familiarity with the queries is helpful in understanding whether two sets of results are different and useful. Queries in both the sets primarily have the informational intent Broder, 2002. The total number of queries was limited by the budget available for the study.

We probed the search engines with the same set of queries at the same time of the day for a period 21 days for the Trends set, and 17 days for the Manual set, during June-July 2014. For Google, we used their *custom search API* [3], and for Bing their *search API*[4]. For both, we recorded the top-k results. The value of k is set to 10 by default, except in the experiments studying the sensitivity of results to the value of k. Every time, we ran the same code from the same machine having the same IP address to minimize noise in the results. Because we were getting the results programmatically through the API, no cookies were used and there was no browser information used by Google or Bing in producing the results Hannak *et al.*, 2013.

[3] code.google.com/apis/console

[4] datamarket.azure.com/dataset/bing/ search

4.2 Representation of Search Results

While our methodology is independent of the specific representation of search results, we employ the snippets of the search results provided by the search engines for this purpose. The snippet of a search result embodies the search engine's semantic understanding of the corresponding document with respect to the given query. The users also heavily weigh the snippet in deciding whether to click on a search result Marcos and González-Caro, 2010. The alternative of using URL representation must first address the well-known problems arising from short URLs Antoniades *et al.*, 2011, un-nomalized URLs Lee *et al.*, 2005; Lei *et al.*, 2010, and different URLs with similar text Bar-Yossef *et al.*, 2009. Unfortunately, there is no agreed upon way to address them and the specific algorithms deployed can have large impact on the conclusions. Furthermore, the users rarely decide whether to look at a document based on the URL they see on the search result page Marcos and González-Caro, 2010.

More in detail, for a given result of a particular query, on a given date, we take the bag-of-words representation of the snippet, after eliminating stopwords. Subsequently, a set of results from a particular search engine, for a given query, is simply the union of the respective bag-of-words representations. For TensorCompare, we keep all words and their frequencies; 0/1 features did not change the trends. For CrossLearnCompare, we keep the top-n words and have binary features. Finally, the distribution of the snippet lengths for Google and Bing was almost identical for all the queries. This ensures a fair comparison between the two engines.

To assess whether snippets are appropriate for comparing the search results, we conducted the following experiment. We inspect the top result given by Google and Bing for a single day, for each of the queries in both Trends and Manual datasets. If for a query, the top result points to the same content, we assign the URL similarity score of 1 to this query, and the score of 0 otherwise. We then compute the cosine similarity between the bag-of-word representations of the snippets produced by

Albert Einstein	American Idol	Antibiotics	Ariana Grande
Avicii	Barack Obama	Beyonce	Cristiano Ronaldo
Derek Jeter	Donald Sterling	Floyd Mayweather	Ford Mustang
Frozen	Game of Thrones	Harvard University	Honda
Jay-Z	LeBron James	Lego	Los Angeles Clippers
Martini	Maya Angelou	Miami Heat	Miami Heat
Miley Cyrus	New York City	New York Yankees	Oprah Winfrey
San Antonio Spurs	Skrillex	SpongeBob SquarePants	Tottenham Hotspur F.C.
US Senate			

Table 1: TRENDS queries

Afghanistan	Alternative energy	Athens	Beatles	Beer
Coup	Debt	Disaster	E-cigarettes	Education
Gay marriage	Globalization	Gun control	IMF	iPhone
Iran	Lumia	Malaria	Merkel	Modi
Paris	Polio	Poverty	Rome	Russia
San Francisco	Self-driving car	Syria	Tesla	Ukraine
Veteran affairs	World bank	World cup	Xi Jinping	Yosemite

Table 2: MANUAL queries

the two search engines for the same query. Figure 2 shows the outcome of this experiment. Each point in this figure corresponds to one query and plots the URL and snippet similarity scores for this query. For clarity, the X and Y axes show ranges beyond [0,1].

We see that for most of the queries for which the snippet similarity was low, the results pointed to different documents; on the other hand, when the similarity of snippets is high, the documents are identical. In both TRENDS and MANUAL, there exist some outliers with pointers to identical documents yet dissimilar snippets (e.g. the query `Tesla` in TRENDS and `US Senate` in MANUAL). Yet, overall, Fig. 2 indicates that snippets are good instruments for content comparison.

Note that we do not consider their ordering in our representation of the search results. Instead, we study the sensitivity of our conclusions to the number of top results, including top-1, top-3, and top-5 (in addition to top-10).

4.3 Exploratory Aggregate Analysis

For a quick, aggregate look at the results produced by the two search engines, we show in Fig. 3 the pairwise word frequency distributions as scatterplots; each point corresponds to a word, and its value on either axis is the (normalized) frequency of occurrence in the results of the respective search engine. If the two search engines were consistently outputting identical results for every query, then all points would lie on the 45 degree line, which is not the case here. However, we observe a trend of many terms having similar frequencies.

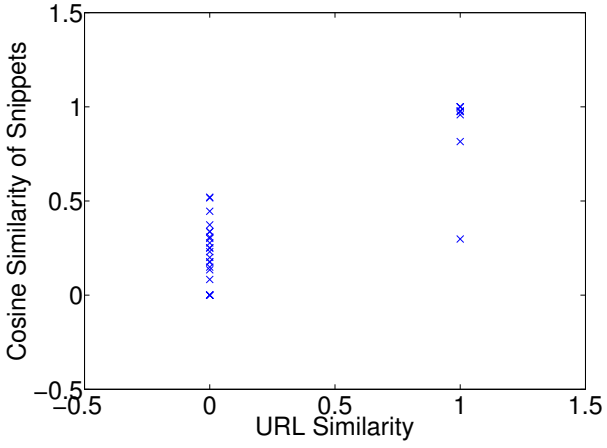
This simple analysis ignores crucial information, such as the query (i.e. the context under which two search engines can be similar or different), as well as the time dimension, thus signifying the need for more specialized analytic tools, such as the proposed methods TENSORCOMPARE and CROSSLEARNCOMPARE.

4.4 Results of TensorCompare

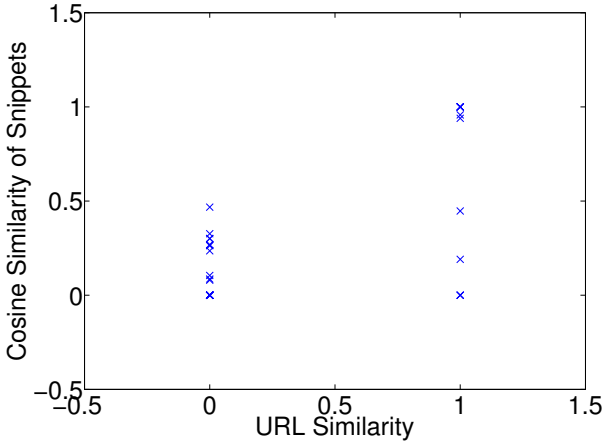
The input tensor to TENSORCOMPARE has modes (query, term, date, search engine). Our data collection results in a $32 \times 36631 \times 21 \times 2$ tensor for the TRENDS dataset and a $35 \times 39725 \times 17 \times 2$ tensor for the MANUAL set. For fitting the PARAFAC decomposition, we use the `CP_APR` algorithm from Chi and Kolda, 2012 that is appropriate for sparse, count data[5]. More specifically, we use Tensor Toolbox from Matlab Bader and Kolda, 2007b, which contains an efficient implementation of this algorithm. The number of components we chose was $R = 20$; however, qualitatively similar behavior was observed for various values for R. As we discussed in Section 3.1, determining the right value for R is a hard problem. To that end, we used the Automatic Relevance Determination heuristic Mørup and Hansen, 2009 to validate our choice. In particular, we computed decompositions for $R = 10$, $R = 15$, $R = 20$, and $R = 25$, and for all those cases, the heuristic indicated that the decomposition was of good quality. The results of TENSORCOMPARE analysis are shown in Figs. 4 and 5. Run-time results are shown in Table 3. Figure 4 shows the similarity of search results, while Fig. 5 shows the temporal profile of each one of the points in Fig. 4.

The first, immediate, observation is that the latent clusters for both query sets behave very similarly. This fact is encourag-

[5]The use of `CP_APR` is shown to perform well for sparse, count data, and is not affected by extremely popular and highly frequent terms. In that sense, it is fair to say that using this algorithm acts effectively as if we used a TF-IDF normalization, which is popular in the Information Retrieval literature.

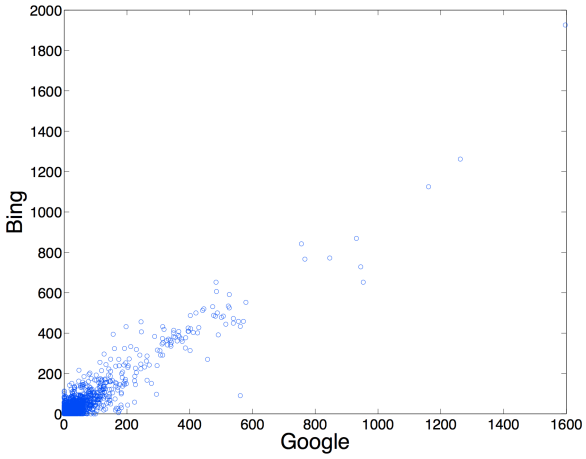


(a) Trends query set

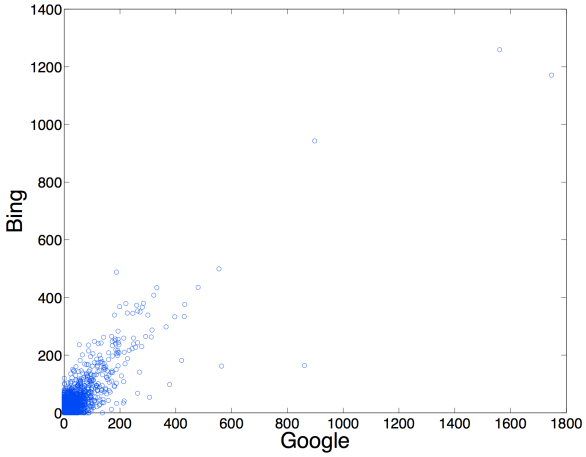


(b) Manual query set

Figure 2: Comparing URL similarity with snippet similarity



(a) Trends query set



(b) Manual query set

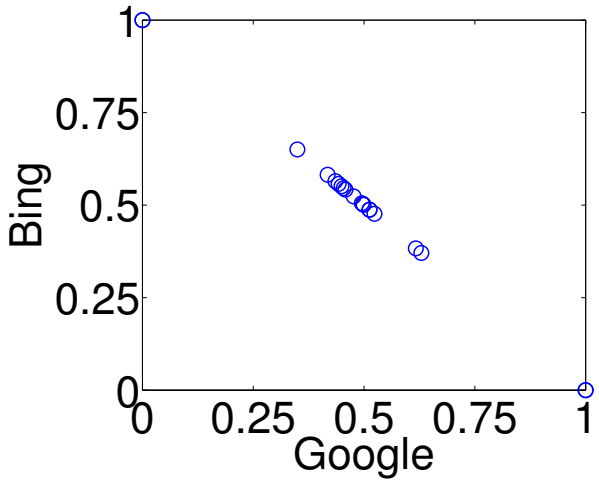
Figure 3: Term distribution between Google and Bing search results.

Dataset	Run-time (sec)
Trends-top10	65.8747 ± 67.9602
Manual-top10	43.4841 ± 29.0559
Trends-top5	37.4425 ± 47.5430
Manual-top5	81.8042 ± 69.1532
Trends-top1	2.1430 ± 1.2636
Manual-top1	4.2782 ± 3.1452

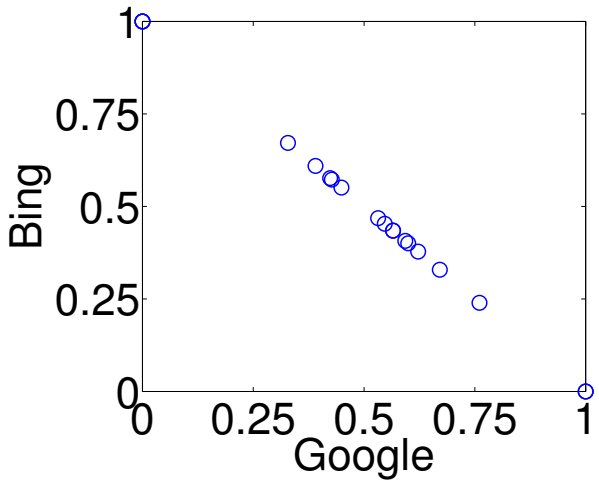
Table 3: Run-times for `CP_APR` for $R = 20$ components, for all the datasets we analyze. We observe that as we reduce the number of results (which reduces the number of non-zero elements in the tensor), the decomposition is faster, which reflects the dependence of the complexity on the number of entries in the data.

ing because it shows that our analysis can be applied to both head and trunk queries. In order to interpret the aforementioned plots, we consult Fig. 1. We observe that Google and Bing produce similar results. This is indicated by the fact that in Fig. 4, the majority of the points lie around the $(0.5, 0.5)$ point (we remind the reader that this point indicates almost exact similarity for the case of two search engines), showing near equal participation of Google and Bing to the majority of the latent clusters. This finding is quite surprising and is in sharp contrast with the past studies. We further observe that that there are somewhat more results unique to Google than Bing since there are more clusters where Google has single participation.

Finally, with respect to the temporal variation of the results, as indicated by Fig. 5, the temporal profile of each cluster is almost uniform across time. This, consequently, means that for both search engines, either in cases where they agree or in cases where they produce somewhat distinct results, their behavior is stable over time, at least as observed during the duration of our study.

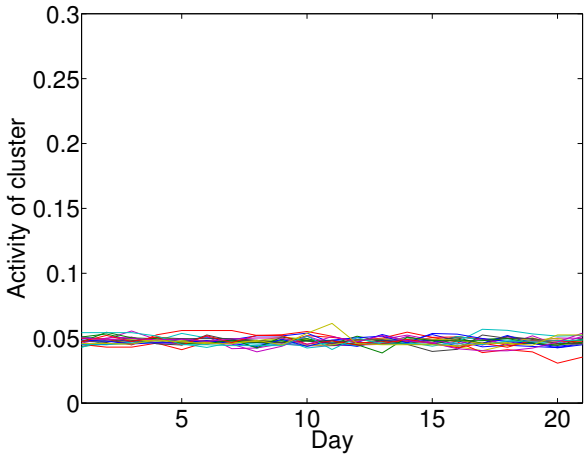


(a) Trends query set

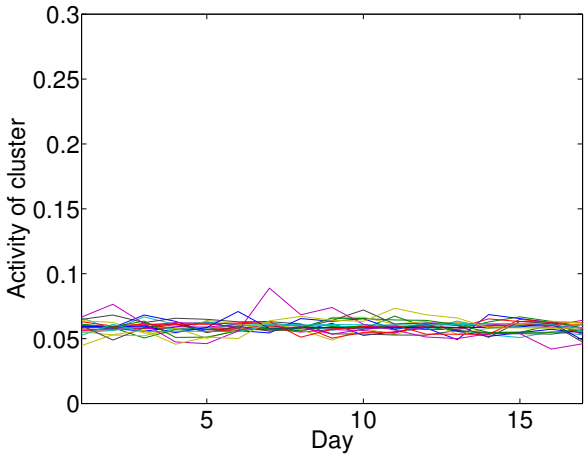


(b) Manual query set

Figure 4: Visualization of TensorCompare for Google and Bing. Values on the x-axis correspond to the membership of Google to a cluster, and values on the y-axis correspond to the membership of Bing. Thus, an (x, y) point on this plot represents one of the clusters of TensorCompare. The closer the points are to the 45-degree line, the more similar are the two search engines.



(a) Trends query set



(b) Manual query set

Figure 5: Temporal profile of latent clusters given by TensorCompare, for Google and Bing. The y-axis corresponds to the membership of the particular day to the cluster of interest. For both query sets, the temporal profile of all clusters is approximately constant over time. In particular, each value for Trends is $\approx 1/21$ and for Manual it is $\approx 1/17$. As stated in Section 3.1, this indicates that both Bing and Google returned persistent results, at least during the duration of our experiment. Due to this uniformity, we overplot all clusters, without making any distinctions.

4.5 Results of CrossLearnCompare

We next present our analysis of the application of CrossLearnCompare to the search results of two engines. To obtain feature space for our instances, we remove terms that are verbatim equal to or contain the query string and then take the 100 highest frequency words for each search engine. We use the union of these two bags of words as the feature space of the training and testing instances. Each such instance is, thus, the vector space representation of a result for a given date and position in the result-set. We use a binary representation, where 1 indicates that the corresponding word appears in the particular instance.

We train one-vs-all linear SVM classifiers for each query set, for each search engine. The performance of the two classifiers of CrossLearnCompare for the two query sets is shown in Fig. 6; the measure of performance used is the standard Receiver Operating Characteristic (ROC) curve Brown and Davis, 2006. There are four curves on the same figure, showing the performance of predicting Bing using Google and vice versa, and for query sets Trends and Manual. Table 4 contains the Area Under the Curve (AUC) for the ROC curves shown in Fig. 6.

Firstly, we observe that the search results are mutually highly predictable for the Trends query set. This implies that

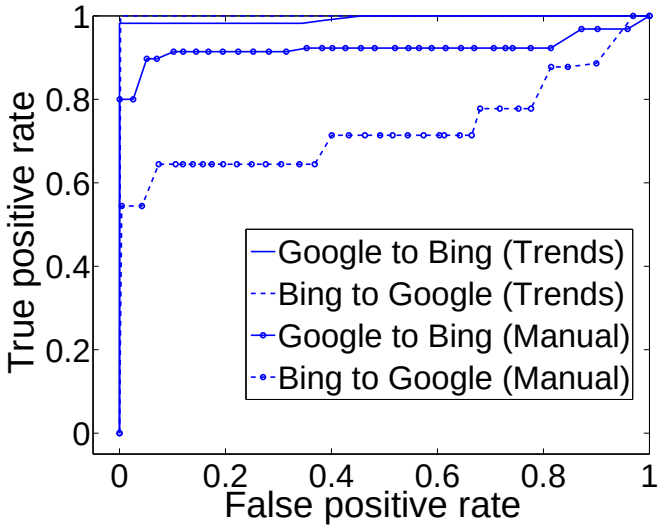


Figure 6: ROC curves produced by CrossLearnCompare (higher is better in terms of classification accuracy). If two search engines were completely mutually predictable, the ROC curve would be exactly on the $(0,0)-(0,1)$ and $(0,1)-(1,0)$ lines. Conversely, if two search engines were completely mutually unpredictable, the ROC curve would lie on the $(0,0)-(1,0)$ and $(1,0)-(1,1)$ lines. Finally, when the classifier is random, the curve would lie on the 45-degree line.

	Trends →	Trends ←	Manual →	Manual ←
Google- Bing	0.99	1.00	0.92	0.73

Table 4: Area Under the Curve (AUC) results for CrossLearnCompare. The right arrow → indicates that we use the left search engine to predict the right one, and ← the converse.

the top results for these popular queries for Google and Bing are very similar. The same behavior continues to be observed for the Manual query set, albeit Google results are somewhat less predictable from Bing results.

4.6 Query Level Analysis

We next present the findings of our overlap analysis at a finer level. TensorCompare yields groups of queries, terms, and search engines, wherein one can study the context under which the search engines agree and disagree. Table 5 shows six exemplar groups where Google and Bing have equal participation; for each group, we show the top-10 snippet terms that were in common within the search results. On the other hand, Table 6 shows exemplar groups where Google and Bing produced distinct clusters. We note that Groups #2 and #3 from the latter table are interesting because they share two of the queries (Afghanistan and Syria). That these queries resulted in clusters exclusive to Google and Bing indicates that the search engine produced different results for them.

CrossLearnCompare naturally yields a query level analysis as it frames the comparison as a multi-class classification problem, where queries are the different classes. Figure 7 shows the mutual predictability of Google and Bing per query, for both query sets. We observe that for Trends, both Google and Bing are highly mutually predictable (therefore have high overlap in search results) for the vast majority of the queries; for Manual, we observe a small decrease in the overlap, nevertheless a large fraction of the queries exhibit high overlap. Consistent with the results of Table 6, we see that Afghanistan and Syria have low overlap, compared to the majority of the queries.

4.7 Validation Using the TRM Method

Recall our discussion of the TRM method Teevan *et al.*, 2011, provided in Section 2. Since the TensorCompare method can be seen as a topic model over the results, the queries, and the search engines, we apply the TRM method to the topics (sets of terms) emerging from TensorCompare. We first apply tensor analysis to the Google and Bing results to obtain their representations in the latent space. We then compute the centroids for the Google and the Bing results topics, and for every result from Google and Bing (for all queries and days), we compute its cosine distance from each centroid. While calculating the centroids, we ignore topics that are shared between Google and Bing and keep those that lie on the $(0,1)$ and $(1,0)$ points of the TensorCompare plots. Essentially, by doing this, we are calculating the largest distance between Google and Bing topics. We present the results of this experiment in Table 7.

		To Google centroid	To Bing centroid
Trends	From Google result	0.13	0.11
	From Bing result	0.11	0.13
		To Google centroid	To Bing centroid
Manual	From Google result	0.20	0.16
	From Bing result	0.13	0.16

Table 7: Similarity from centroids

For the Trends set, the difference between the four distances is not statistically significant, whereas for the Manual the differences are statistically significant, albeit very small. In general, Table 7 shows the distance of a result from both the Google and the Bing topic is small, corroborating our observation that the overlap in results among Google and Bing is large.

4.8 Sensitivity Analysis

One might wonder how sensitive are our conclusions to the fact that we analyzed the top-10 search results. To this end, we apply TensorCompare and CrossLearnCompare to the top-5, top-3, and top-1 search results, for both Trends and Manual query sets. Figures 8 and 9 show the results of this analysis for top-5 and top-1 for TensorCompare and CrossLearnCompare respectively. The results for top-3 lie between top-5 and top-1 and have been omitted.

We see that our earlier findings are robust and consistent with the ones presented here. A few specific remarks follow:

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Queries	San Antonio Spurs Miami Heat	Donald Sterling Los Angeles Clippers	Albert Einstein	self-driving car	Athens San Francisco	World bank Modi
Terms	news miami heat san antonio spurs scores stats team schedule	sterling donald clippers angeles owner 2014 nba racist former team	einstein albert physicist born biography 1879 march quotes german web	google driving car self cars autonomous steve announced own steering	information san francisco city county offers official guide greece services	world bank narendra minister prime india countries development international loans

Table 5: Query groups where Google and Bing have equal participation (TensorCompare).

	Group 1	Group 2	Group 3	Group 4
Queries	Syria education Afghanistan Merkel	Frozen American Idol Lego	San Francisco Rome Athens Russia	Afghanistan Iran Syria
Terms	city breaking videos world com information people photos politics angela	news american digita com disney blu ray 2013 dvd season	news hotels san francisco attractions restaurants capital information tours italy	idol country economyl politics information world republic breaking islamic iraq
Search Engine	Google	Google	Bing	Bing

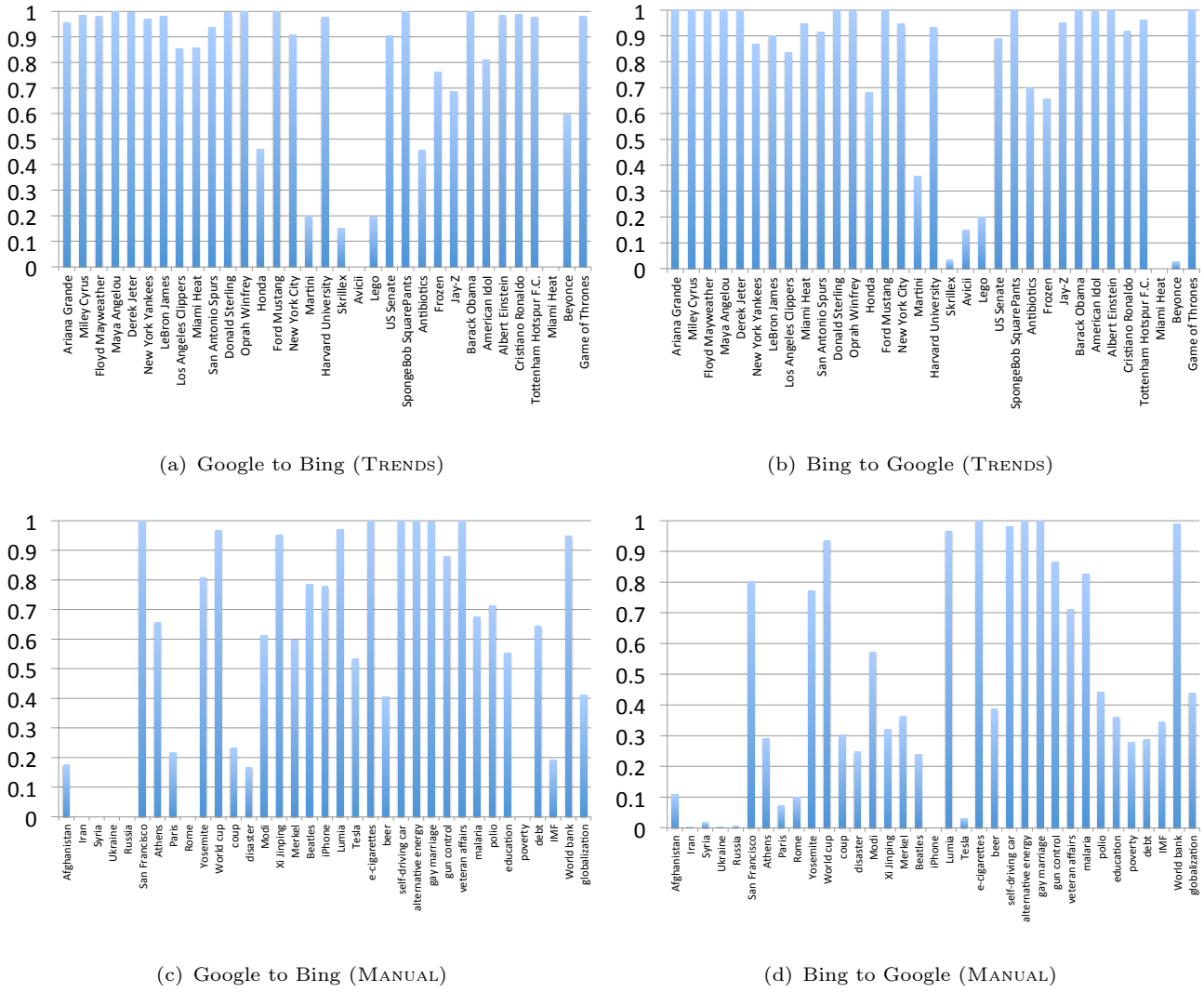
Table 6: Query groups unique to Google or Bing (TensorCompare).

- The two search engines continue to exhibit more similar results for the Trends query set (head queries) than the Manual set (trunk queries).
- Using top-5 as the cut-off, the similarity is slightly higher than using top-10. This indicates that it is more likely that the search engines will have an exclusive result below position 5.
- For the single top result, even though there is similarity, the top result is not necessarily the same (but the manual inspection reveals that the top result of one is almost always present in the top-5 of the other).
- The results of CrossLearnCompare reinforce the findings of TensorCompare. For top-5, the classifier learned using the results of one search engine is able to quite accurately predict the results of the other. However, given the sensitivity of the top result to the position, the performance degrades for top-1.

5 Summary, Limitations, and Future Work

We introduced two novel tools for studying the similarity and distinctiveness of web results of search engines. Our main observation, stemming from our analysis, is that Google and Bing exhibited a significant degree of similarity in the semantics of their search results in our data set. This observation is in sharp contrast to the prior published work where minimal overlap is reported. A fair interpretation of our observation is stating that the visual experience of users in Google and Bing is very similar for the queries we studied. This can be seen as an upper bound to the exact number of overlapping results, which the prior work is trying to estimate.

Our results depend on the particular choice of queries. It is possible that if one adversary chose a set of queries from the long tail, our tools would indicate minimal overlap, however this would not reflect realistic users' search patterns. Our selection of queries, within the budget limitation of the study, strikes a balance between queries on the head of the distribution (Trends) and queries that span a wider spectrum of



(a) Google to Bing (TRENDS)

(b) Bing to Google (TRENDS)

(c) Google to Bing (MANUAL)

(d) Bing to Google (MANUAL)

Figure 7: Prediction precision per query (CROSSLEARNCOMPARE). For the majority of queries, Google and Bing are highly mutually predictable.

popularity (MANUAL). Furthermore, our results rely on the assumption that the snippet similarity is a good proxy for web-page similarity. We experimentally validate this assumption for the queries that we studied, however, an adversarially chosen set of queries (e.g., on the long tail, as before) or a set of adversarial snippet generators (e.g., choosing terms at random from the web-page), may violate this assumption.

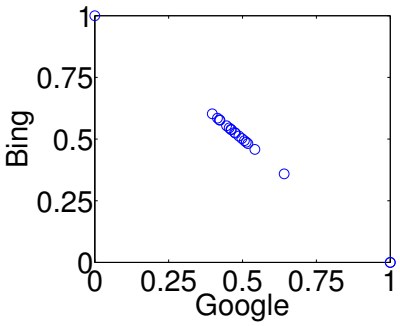
We can only speculate why there is greater convergence in the results produced by the two search engines. They include deployment of greater amount of resources by search engines to cover a larger fraction of the indexable Web, much more universal understanding of search engine technologies, and the use of similar features in ranking the search results.

In the future, we would like to explore the feasibility of building a search engine that uses signals not used by Google and Bing and yet produces useful results. Such signals could possibly come from social networks. It will also be interesting to explore the practicality of explicitly designing-in diversity into search engines Maltese *et al.*, 2009.

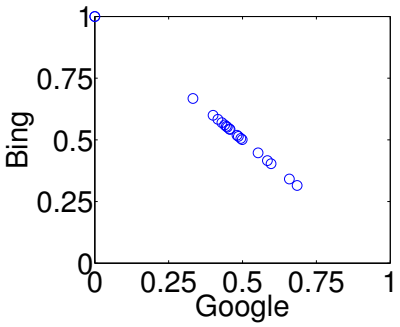
Proof of Lemma 1

Proof. Consider a tensor $\underline{\mathbf{X}}$ with dimensions $I \times J \times 2$ (in our case, the first mode corresponds to queries, the second to results, and the third to search engines). Assume that $\underline{\mathbf{X}}$ is rank one, which means that there is one component in its PARAFAC decomposition. In the frontal slice corresponding to the first search engine (Slice 1 in Fig. 10), we have Q queries and T results forming a perfect block, which we assume to be filled with 1's. The second slice, which corresponds to the second search engine, has a block that spans only a fraction of the queries and results of Slice 1.

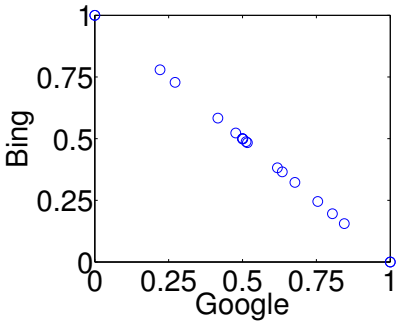
We assume that the components $\mathbf{a}, \mathbf{b}$ of the PARAFAC decomposition are normalized by their ℓ_2 norm, and the scaling is absorbed in $\mathbf{c}$. We further assume that the components are



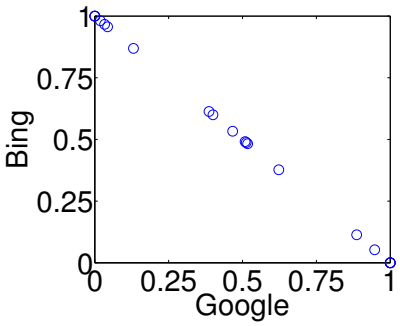
(a) TRENDS top-5



(b) MANUAL top-5

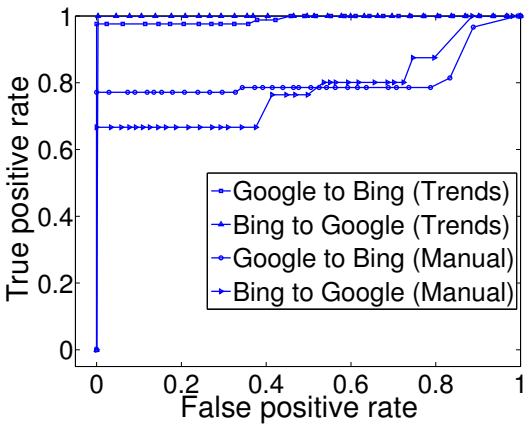


(c) TRENDS top-1

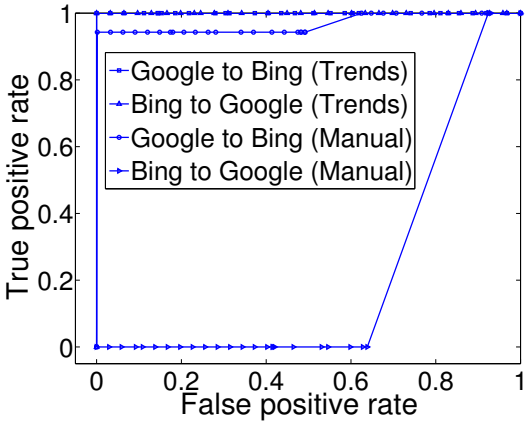


(d) MANUAL top-1

Figure 8: TENSORCOMPARE sensitivity



(a) top-5



(b) top-1

Figure 9: CROSSLEARNCOMPARE sensitivity

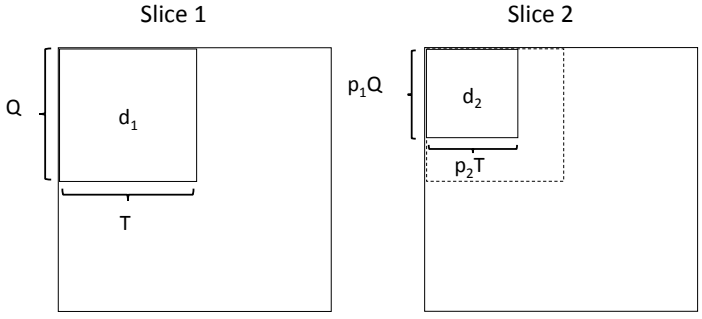


Figure 10: The two slices of $\underline{\mathbf{X}}$.

non-negative.

Let $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{c}}$ be the optimal solution. An upper bound $\mathbf{a}, \mathbf{b}, \mathbf{c}$ to the optimal is the following: The first Q elements of $\mathbf{a}$ will be equal to $\frac{1}{\sqrt{Q}}$ (the rest are zero), and the first T elements of $\mathbf{b}$ will equal $\frac{1}{\sqrt{T}}$. This implies that the coefficients of $\mathbf{c} = \begin{bmatrix} c_1 & c_2 \end{bmatrix}^2$, which multiply $\mathbf{a}\mathbf{b}^T$ in order to approximate the respective slices of $\underline{\mathbf{X}}$, will be proportional to the respective densities of the blocks in either slice, i.e. $c_1 \propto d_1$ and $c_2 \propto d_2$ Making this uniformity assumption for the non-zero elements

of $\mathbf{a}, \mathbf{b}$ allows us to bound the ratio of the coefficients of $\hat{\mathbf{c}}$ by the ratio of the densities of the blocks in each slice. More specifically, we have

$$\frac{\hat{c}_1}{\hat{c}_2} \leq \frac{d_1}{d_2} = \frac{QT}{p_1 Q p_2 T} = \frac{1}{p_1 p_2}.$$

Hence, $\hat{c}_2 \leq p_1 p_2 \hat{c}_1$. If we substitute $y = \hat{c}_2$ and $x = \hat{c}_1$, as they correspond in Fig. 1, then we have shown the desired upper bound. □

References

Antoniades, D., I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis. "we.b: The web of short URLs". In: *20th international conference on World Wide Web*. ACM. 715–724.

Bader, B. W. and T. G. Kolda. "Efficient MATLAB computations with sparse and factored tensors". *SIAM Journal on Scientific Computing*. 30(1): 205–231.

Bader, B. W. and T. G. Kolda. "Matlab tensor toolbox version 2.2". *Albuquerque, NM, USA: Sandia National Laboratories*.

Bar-Ilan, J. "Search engine ability to cope with the changing Web". In: *Web dynamics*. Springer. 195–215.

Bar-Yossef, Z., I. Keidar, and U. Schonfeld. "Do not crawl in the DUST: different urls with similar text". *ACM Transactions on the Web*. 3(1): 3.

Bharat, K. and A. Broder. "A technique for measuring the relative size and overlap of public web search engines". *Computer Networks and ISDN Systems*. 30(1): 379–388.

Blei, D. M., A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation". *the Journal of machine Learning research*. 3: 993–1022.

Broder, A. "A taxonomy of web search". *ACM Sigir forum*. 36(2): 3–10.

Bro, R. and H. A. Kiers. "A new efficient method for determining the number of components in PARAFAC models". *Journal of chemometrics*. 17(5): 274–286.

Brown, C. D. and H. T. Davis. "Receiver operating characteristics curves and related decision measures: A tutorial". *Chemometrics and Intelligent Laboratory Systems*. 80(1): 24–38.

Chi, E. C. and T. G. Kolda. "On tensors, sparsity, and nonnegative factorizations". *SIAM Journal on Matrix Analysis and Applications*. 33(4): 1272–1299.

Chu, H. and M. Rosenthal. "Search engines for the World Wide Web: A comparative study and evaluation methodology". In: *American Society for Information Science*. Vol. 33. 127–135.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. "Indexing by latent semantic analysis". *JASIS*. 41(6): 391–407.

Ding, W. and G. Marchionini. "A Comparative Study of Web Search Service Performance". In: *ASIS Annual Meeting*. Vol. 33. ERIC. 136–42.

Enge, E., S. Spencer, J. Stricchiola, and R. Fishkin. *The art of SEO*. O'Reilly.

Federal Communications Commission. "Editorializing by Broadcast Licensees". Washington, DC: GPO.

Gauch, S. and G. Wang. "Information Fusion with ProFusion". In: *1st World Conference of the Web Society*.

Guan, Z. and E. Cutrell. "An eye tracking study of the effect of target rank on web search". In: *SIGCHI conference on Human factors in computing systems*. ACM. 417–420.

Gulli, A. and A. Signorini. "The indexable web is more than 11.5 billion pages". In: *14th international conference on World Wide Web*. ACM. 902–903.

Hannak, A., P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. "Measuring personalization of web search". In: *22nd international conference on World Wide Web*. ACM. 527–538.

Harshman, R. A. "Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis". *Tech. rep.* UCLA.

Håstad, J. "Tensor rank is NP-complete". *Journal of Algorithms*. 11(4): 644–654.

Kang, U., E. Papalexakis, A. Harpale, and C. Faloutsos. "Gigatensor: scaling tensor analysis up by 100 times - algorithms and discoveries". In: *18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 316–324.

Kolda, T. G. and B. W. Bader. "Tensor decompositions and applications". *SIAM review*. 51(3): 455–500.

Lancaster, F. W. and E. G. Fayen. *Information Retrieval On-Line*. Melville Publishing Co.

Lawrence, S. and C. L. Giles. "Searching the world wide web". *Science*. 280(5360): 98–100.

Lee, S. H., S. J. Kim, and S. H. Hong. "On URL normalization". In: *Computational Science and Its Applications–ICCSA 2005*. Springer. 1076–1085.

Lei, T., R. Cai, J.-M. Yang, Y. Ke, X. Fan, and L. Zhang. "A pattern tree-based approach to learning URL normalization rules". In: *19th international conference on World Wide Web*. ACM. 611–620.

Lewandowski, D. *Web search engine research*. Emerald Group Publishing.

Maltese, V., F. Giunchiglia, K. Denecke, P. Lewis, C. Wallner, A. Baldry, and D. Madalli. *On the interdisciplinary foundations of diversity*. University of Trento.

Marcos, M.-C. and C. González-Caro. "Comportamiento de los usuarios en la página de resultados de los buscadores. Un estudio basado en eye tracking". *El profesional de la información*. 19(4): 348–358.

Mørup, M. and L. K. Hansen. "Automatic relevance determination for multi-way models". *Journal of Chemometrics*. 23(7-8): 352–363.

Papalexakis, E. E., U. Kang, C. Faloutsos, N. D. Sidiropoulos, and A. Harpale. "Large Scale Tensor Decompositions: Algorithmic Developments and Applications." *IEEE Data Eng. Bull.* 36(3): 59–66.

Pirkola, A. "The Effectiveness of Web Search Engines to Index New Sites from Different Countries". *Information Research: An International Electronic Journal*. 14(2).

Purcell, K., J. Brenner, and L. Rainie. *Search engine use 2012*. Pew Internet & American Life Project.

Selberg, E. and O. Etzioni. "Multi-service search and comparison using the MetaCrawler". In: *4th international conference on World Wide Web*.

Sidiropoulos, N., E. E. Papalexakis, and C. Faloutsos. "Parallel randomly compressed cubes: A scalable distributed architecture for big tensor decomposition". *Signal Processing Magazine, IEEE*. 31(5): 57–70.

Spink, A., B. J. Jansen, and C. Wang. "Comparison of major Web search engine overlap: 2005 and 2007". In: *14th Australasian World Wide Web Conference*.

Spink, A., B. J. Jansen, C. Blakely, and S. Koshman. "A study of results overlap and uniqueness among major web search engines". *Information Processing & Management*. 42(5): 1379–1391.

Stroud, N. J. and A. Muddiman. "Exposure to News and Diverse Views in the Internet Age". *ISJLP*. 8: 605.

Teevan, J., D. Ramage, and M. R. Morris. "# TwitterSearch: a comparison of microblog search and web search". In: *4th ACM international conference on Web search and data mining*. ACM. 35–44.

Wilkinson, D. and M. Thelwall. "Search markets and search results: The case of Bing". *Library & Information Science Research*. 35(4): 318–325.