Whither Social Networks for Web Search?

Rakesh Agrawal∗
Data Insights Laboratories
ragrawal@acm.org

Behzad Golshan∗
Boston University
behzad@bu.edu

Evangelos Papalexakis∗
Carnegie Mellon University
epapalex@cs.cmu.edu

ABSTRACT

Access to diverse perspectives is essential for inculcating and nurturing an informed citizenry. Google and Bing have emerged as the duopoly that largely arbitrates what English language documents are seen by Web searchers. A recent study shows that there is now a large overlap in the top organic search results produced by them. Thus, citizens may no longer be able to garnish different perspectives by probing different search engines.

We present the results of our empirical study that indicates that by mining Twitter data one can obtain search results that are quite distinct from those produced by Google and Bing. Additionally, the users found those results to be quite informative in our user study. The gauntlet is now on search engines to test whether our findings hold in their infrastructure for different social networks and whether enabling diversity has sufficient business imperative for them.

1. INTRODUCTION

The fairness doctrine contends that citizens should have access to diverse perspectives as exposure to different views is beneficial for the advancement of humanity [22]. The World Wide Web is now widely recognized as the universal information source. Without question, content representing diverse perspectives exist on the Web, almost on any topic. However, this does not automatically ensure that citizens encounter them [50].

Search engines have become the primary tool used to access the Web content[42]. In particular, it is the duopoly of Google and Bing that largely arbitrates what documents people see, especially from the English language Web (Yahoo’s Web search is currently powered by Bing).

A recent study [2] indicates that there is now a large overlap in the top-10 organic search results produced by Google and Bing. These are the results that get almost all of the clicks as users rarely look at results at lower positions [23] [28]. This overlap was found to be even more pronounced in the top-5 results and the results of queries in which citizens exhibited large interest. These findings are in sharp contrast to the past studies that found little overlap in the results of the then popular search engines. The implication is that citizens may no longer be able to garnish different perspectives by obtaining results for the same query from different search engines.

This paper investigates whether data mining of social networks can help Web search engines imbue their search results with useful diversity [57]. Specifically, we present the results obtained by mining the real-life Twitter data that demonstrate:

1. We are able to obtain search results, even by simply analyzing the retweet graph, which are quite distinct from the Web results for the same query.

2. The users have judged those results to be quite informative.

Our fond hope is that this paper will provide impetus to the commercial search engines to examine the applicability of our findings in their infrastructure. We used Twitter in our case study because it is still possible to selectively crawl Twitter. Our secondary hope is that our findings will instigate other social networks to make their data available to data mining researchers for exploring the use of social networks to obtain new signals for enhancing the Web search (or building new search engines!).

Paper Layout.

The structure of the rest of the paper is as follows. We begin by discussing related work in Section 2. In Section 3, we describe the data mining tools we employed for conducting our study. Section 4 presents the results of our analysis using data from Google, Bing, and Twitter. Section 5 presents the user study for assessing the usefulness of our findings. We conclude with a summary and future directions in Section 6.

2. RELATED WORK

Three lines of research are most relevant to our work: i) overlap between the results of the search engines, ii) social search technologies, and iii) integration of social search results into Web search. We review all three in this section. Note that we use the term “social search” to mean searches conducted over databases of socially generated content, although this term often refers broadly to the process of finding information online with the assistance of any number of social resources such as asking others for answers or two people searching together [55].

2.1 Overlap Studies

Since their advent in early 90’s, there has been considerable interest in understanding how distinct are the results produced by the prevalent Web search engines. Ding and Marchionini measured
and observed in 1996 a low level of result overlap between InfoSeek, Lycos, and OpenText. Around the same time, Selberg and Etzioni found that each of Galaxy, InfoSeek, Lycos, OpenText, WebCrawler and Yahoo returned mostly unique results. Also in 1996, Gauch, Wang and Gomez found that a metasearch engine that fused the results of AltaVista, Excite, InfoSeek, Lycos, OpenText, and WebCrawler provided the highest number of relevant results. Bharat and Broder estimated the overlap between the Websites indexed by HotBot, Alta Vista, Excite and InfoSeek in November 1997 to be only 1.4%. Lawrence and Giles, in their study of AltaVista, Excite, HotBot, Infoseek, Lycos, and Northern Light published in 1998, found that the individual engines covered from 3 to 34% of the indexable Web. Spink et al. studied the overlap between the results of four search engines, namely MSN (predecessor of Bing), Google, Yahoo and Ask Jeeves, using data from July 2005. They found 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%. One way the users dealt with low overlap was by manually executing the same query on multiple search engines. Analyzing six months of interaction logs from 2008-2009, White and Dumais found that 72.6% of all users used more than one engine during this period, 50% switched engines within a search session at least once, and 67.6% used different engines for different sessions. Their survey revealed three classes of reasons for this behavior: dissatisfaction with the quality of results in the original engine (dissatisfaction, frustration, expected better results, totaling 57%), the desire to verify or find additional information (coverage/verification, totaling 26%, curiosity), and user preferences (destination preferred, destination typically better, totaling 12%). Another way the problem of low overlap was addressed was by developing metasearch engines (e.g. InFind, MetaCrawler, MetaFerret, ProFusion, SavvySearch). A metasearch engine automatically queries a number of search engines, merges the returned lists of results, and presents the resulting ranked list to the user as the search of the query. Note that with either manual or automated approach, the user ends up seeing multiple perspectives.

A recent study, using data from June-July 2014, however, found large overlap between the top-10 search results produced by Google and Bing. This overlap was found to be even more pronounced in the top-5 results and the results of head queries. Some plausible reasons for greater convergence in the search results include deployment of greater amount of resources by search engines to cover a larger fraction of indexable Web, much more universal understanding of search engine technologies, and the use of similar features in ranking the search results. A consequence of this convergence is that it becomes now harder for people to access diverse perspectives.

Contrary to the rich literature on overlap between the results produced by the Web search engines, the only prior work we could find on overlap between Web and Social search results appears in Section 5 of TRM Study. They extracted snippets of all search results from Bing search logs for 42 most popular queries for one week in November 2009. They 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). They found that the average similarity of Twitter posts to the Twitter centroid was higher than the Web results’ similarity to the Web centroid. The issue of usefulness of Twitter results is not addressed in their paper.

We shall see that our study considers head as well as trunk queries and encompasses both Google and Bing. We also employ different data mining tools in our study. Specifically, our TensorCompare uses tensor analysis to obtain low-dimensional representation of search results since the method of moments for LDA reduces to canonical decomposition of a tensor, for which scalable distributed algorithms exist. Our CrossLearnCompare, uses a novel cross-engine learning to quantify the similarity of snippets and tweets. Additionally, we provide a user study demonstrating the usefulness of the Twitter results. We will have more to say quantitatively about the TRM study when we present our experimental results in Section 4.

2.2 Social Search

In addition to being considered a social media and a social network, Twitter may also be viewed as an information retrieval system that people can utilize to produce and consume information. Twitter today receives more than 500 million tweets per day at the rate of more than 33,000 tweets per second. More than 300 billion tweets have been sent since the founding of Twitter in 2006 and it receives more than 2 billion search queries every day. Twitter serves these queries using an inverted index tuned for real-time search, called EarlyBird, described in [14]. While this search service excels at surfacing breaking news and events in real time and it does indeed incorporate relevance ranking, it is a feature that the system designers themselves consider that they have “only begun to explore”.

The prevailing perception is that much of the content found on Twitter is of low quality and the keyword search as provided by Twitter is not effective. In response, there has been considerable research aimed at designing mechanisms for finding good content from Twitter. In many of the proposed approaches, retweet count alone or in conjunction with textual data, author’s metadata, and propagation information play a prominent role. The intuition is that if a tweet is retweeted multiple times, then several people have taken the time to read it, decide it is worth sharing, and then actually retweeted it, and hence it must be of good quality. But, of course, one needs to remove socialware and other spam before using retweet count. Other approaches include using the presence of a URL as an indicator, link analysis on the follows and retweets graphs, clustering taking into account the size and popularity of a tweet, its audience size, and recency, and the semantic approaches including topic modeling. See overviews in [53, 54, 55] for additional perspectives.

In this work, we are not striving to create the best possible social search engine, but rather investigate whether the results obtained using signals from a social network could be substantially different from a Web search engine and yet useful. Thus, in order to avoid confounding between multiple factors, we shall use a simple social search engine that ranks tweets based on retweet analysis.

2.3 Integration of Web and Social search

One of the problems with Twitter search has been that, while it is easy to discover current tweets and trending topics, it is much more difficult to search over older tweets and determine, say, what the fans were saying about the Seahawks during the 2014 Super Bowl. Beginning November 18, 2014, however, it has become possible to search over the entire corpus of public tweets. Still, our own experiments indicate that the ranking continues to be heavily biased towards recency.
Bing has been including a few tweets related to the current query on its search result page, at least since November 2013. However, it is not obvious for what queries this feature is triggered and what tweets are included. For example, on February 12, 2015 at 1:42AM, our query “Greece ECB” brought only one tweet on Bing’s result page, which was a retweet from Mark Rauffalo from two days ago. Bing also offered a link titled “See more on Twitter” below this tweet. Clicking this link took us to a Twitter page, where the top tweet was from 14 minutes ago with the text “ECB raises pressure on Greece as Tsipras meets EU peers”! Since June 2014, one can also search Bing by hashtag, look up specific Twitter handles, or search for tweets related to a specific celebrity. As we write this paper, Google is said have struck a deal with Twitter in early February that will allow tweets to be shown in Google search results sometime during the first half of this year.

There is also research on how Web search can be improved using signals from Twitter. For example, Rowlands et al. [45] propose that the text around a URL that appears in a tweet may serve to add supplementary terms or add weight to existing terms in the corresponding Web page and that the reputation or authority of the tweeter may serve to weight both annotations and query-independent popularity. Similarly, Dong et al. [18] advocate using Twitter stream for detecting fresh URLs as well as for computing features to rank them. We propose to build our future work upon some of these ideas.

3. DATA MINING TOOLS

We next review the data mining tools for analyzing and comparing search engine results, introduced in [2]. One, called TensorCompare, uses tensor analysis to derive low-dimensional representation of search results. The other, called CrossLearnCompare, uses cross-engine learning to quantify their similarity.

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. These results can be represented in a four mode tensor $X$, where (query, result, time, search engine) are the four modes [31]. A result might be in the form of a set of URLs or a set of keywords where (query, result, time, search engine) are the four modes [31].

A result might be in the form of a set of URLs or a set of keywords representing the corresponding pages. The tensor might be binary valued or real valued (indicating, for instance, frequencies).

This tensor can be analyzed using the so-called canonical or PARAFAC decomposition [27], which decomposes the tensor into a sum of rank-one tensors:

$$X \approx \sum_{r=1}^{R} \lambda_r \circ b_r \circ c_r \circ d_r,$$

where the $(i,j,k,l)$-th element of $a \circ b \circ c \circ d$ is simply $a(i)b(j)c(k)d(l)$.

The vectors $a_r$, $b_r$, $c_r$, $d_r$ are usually normalized, with their scaling absorbed in $\lambda_r$. For compactness, the decomposition is represented as matrices $A,B,C,D$. The decomposition of $X$ to $A,B,C,D$ gives a low rank embedding of queries, results, timings, and search engines respectively, corresponding to the aforementioned clusters. This implies that we are able to track the temporal behavior of a cluster of semantically similar search engines for a set of queries.

The factor matrix $D$ 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 $D$ being the cluster indicator matrix: the $(i,j)$ entry of $D$ 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 $D$. 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 between search engines.

Fig. 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) shows 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, multiple points are shown along the diagonal for the sake of generality.

Fig. 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. For generality, multiple points are shown on each axis.

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

TensorCompare also allows the tracking of 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 $C$ 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 $1/T$. 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 day profile will have the value approximately equal to one corresponding to that day, and approximately zero for the rest of the days.
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 the classifier to perform in a way that it is correlated with the degree of similarity between A and B.

One can get 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.

4. EVALUATION

We next present the results of the empirical evaluation we performed using data from Google, Bing, and Twitter.

4.1 Social Pulse

For concreteness, we first specify a simple social search engine, which we shall henceforth refer to as Social Pulse. We are not striving to create the best possible search engine, but rather investigate whether the results obtained using signals from a social network could be substantially different from a Web search engine and yet useful. Thus, instead of employing a large set of features (see Section 2.2), we purposefully base the Social Pulse’s ranker on one single feature in order to be able to make sharp conclusions and to avoid confounding between multiple factors.

Social Pulse uses Twitter as the social medium. For a given query, Social Pulse first retrieves all tweets that pertain to that query. Multiple techniques are available in the literature for this purpose (e.g. [9,10,11,31]). We choose to employ the simple technique of checking for the presence of the query string in the tweet. Subsequently, Social Pulse ranks the retrieved tweets with respect to the number of re-tweets (more precisely, the number of occurrences of the exact same tweet without having necessarily been formally re-tweeted).

Arguably, one can restrict the attention to only those tweets that contain at least one URL [3]. However, we have empirically observed that highly re-tweeted tweets, in spite of containing no URL, usually provide high quality result. Hence, Social Pulse uses these tweets as well.

4.2 Data Set

We conducted the study for the same two sets of queries that were used in the study of distinctiveness in the Web search results [2]. The TRENDS set (Table 1) contains the most popular search terms from different categories from Google Trends during April 2014. These are referred to as head queries. The MANUAL set (Table 2) consists of hand-picked queries by the authors, referred 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 very helpful in understanding whether two sets of results are different and useful. Queries in both the sets primarily have the informational intent [12]. The total number of queries was limited by the budget available for the study.

We probed the search engines with identical queries at the same date of the year for a period of 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 (code.google.com/apis/console), and for Bing their search API (datamarket.azure.com/dataset/bing/search). Twitter data consists of 1% sample of tweets obtained using Twitter API. We ran the same code from the same machine having the same IP address to minimize noise in the results [20]. In all cases, 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.

As pointed out in [2], the URL representation of search results suffers from the problems arising from short URLs [7, un-nomalized URLs [35, 36], and different URLs with similar text [8]. We again use, therefore, the vector space representation of the results shown to the users on the search result page. Drawing upon the eyetracking study that the users decide whether to click on a result primarily based on the snippet [35], we use the snippet of a search result for computing its representation. In the case of Social Pulse, the text of a tweet is treated as snippet for this purpose. Snippets and tweet texts respectively have also been used in the study of overlap between the results of Web search and social search in [53].

For a given result of a particular query, on a given date, we take a bag-of-words representation of the snippet, after eliminating stop-words. 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. Finally, we note that the distribution of the snippet lengths for Google, Bing, and Social Pulse was almost identical for all the queries that we tested. This ensures a fair comparison between the engines.

4.3 Data Mining

Our data collection resulted 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 implementing TENSORCOMPARE, we use the Tensor Toolbox for Matlab [7], using KL-Divergence as the loss function. The number of components for the decomposition was $R = 25$; however, qualitatively similar behavior was observed for various values of $R$. For CROSSLEARNCOMPARE, the feature space is obtained by removing terms from search results that are verbatim equal to or contain the query string and then taking the top 100 terms for each search engine. We use the union of these two bags of words as the feature space of the training and testing instances. We use a binary representation, where 1 indicates that the corresponding term appears in the particular instance. We train one-vs-all linear SVM classifiers for each query set, for each search engine. The measure
of performance used is the standard Receiver Operating Characteristic (ROC) curve [13].

4.4 Findings

We present the results of comparing search results of Social Pulse first to that of Google and then Bing.

![Diagram: AUC results for CrossLearnCompare comparing Google, Social Pulse, and Bing.](image)

**Figure 2:** Social Pulse vs. Google

**Figure 3:** Social Pulse vs. Bing

4.4.1 Social Pulse versus Google

Figures 2(a) through 2(d) and Table 3 show the results. Recalling the guide for interpreting TensorCompare plots in Fig. 1, we see from Figs. 2(a) 2(b)

1. There exists a number of results exclusive to either search engine as indicated by multiple points around (0, 1) and (1, 0).

2. For the non-exclusive results, the points are not concentrated on (0.5, 0.5) (which would have indicated similar results), but are rather spread out.

This suggests that Social Pulse and Google provide distinctive results to a great extent.

For the Trends dataset in Fig. 2(a) there is a cloud of clusters around (0.7, 0.3), which indicates that Google has greater participation in these results than Social Pulse. Figure 2(c) and the Area Under the Curve (AUC) results in Table 3 also show that using Google to predict Social Pulse works relatively better than the converse for this dataset. This asymmetry suggests that the Twitter engines for non-head queries.

We repeated the preceding analysis, but by using Bing search results rather than Google this time. Figures 3(a) and 3(b) show the results for TensorCompare for Trends and Manual respectively, and Figs. 3(c) and 3(d) for CrossLearnCompare. Table 4 shows the AUC results for CrossLearnCompare. These results are qualitatively similar those obtained using Google search results, which is not a surprise given the finding in [2] that Google and Bing have significant overlap in their search results. However, this sensitivity analysis employing another commercial search engine further reinforces the conclusion that Social search can yield results quite different from the ones produced by the conventional Web search.

4.4.2 Social Pulse versus Bing

![Diagram: AUC results for CrossLearnCompare comparing Bing and Social Pulse.](image)

**Table 3:** Area Under the Curve (AUC) results for CrossLearnCompare comparing Google and Social Pulse.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>0.6592</td>
<td>0.6449</td>
<td>0.4149</td>
<td>0.7844</td>
</tr>
</tbody>
</table>

- 0.2
- 0.4
- 0.6
- 0.8
- 1

<table>
<thead>
<tr>
<th>Bing</th>
<th>Social Pulse</th>
<th>Manual</th>
<th>Google</th>
<th>Social Pulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.8065</td>
<td>0.6308</td>
<td>0.4388</td>
<td>0.8284</td>
</tr>
</tbody>
</table>

**Table 4:** Area Under the Curve (AUC) results for CrossLearnCompare comparing Bing and Social Pulse.

2. For the non-exclusive results, the points are not concentrated on (0.5, 0.5) (which would have indicated similar results), but are rather spread out.

In contrast, for the Manual dataset in Fig. 3(a) the non-exclusive points are relatively more dispersed along the line that connects (0, 1) and (1, 0) and there are clusters in which Social Pulse is more prominent. We also find that now predicting Google using Social Pulse works better than the converse (Figs. 3(c) and 3(d)). Collectively, they quantitatively validate the intuition that social networks might have content very different from that indexed by Web search engines for non-head queries.

4.4.3 Query Level Analysis

In order to gain further insight into mutual predictability of Web and Social search, we looked at three queries that have the highest and lowest predictability for each search engine and query set, when using CrossLearnCompare analysis. Tables 5 and 6 show the results with respect to Google; the insights gained were similar for Bing.

We see that the timely queries, like World cup or gay marriage, have high mutual predictability. Indeed, timeliness creates relevance; the same information gets retweeted and clicked a lot. Queries like Maya Angelou and Albert Einstein are also highly mutually
predictable, in part because people tend to tweet quotes by them, which tend to surface to Web search results as well.

On the other hand, queries such as globalization and poverty have low predictability. These queries are informational queries with large scope. However, it seems that the content people retweet a lot for these queries is not the same as what is considered authoritative by the Web search ranking algorithms. We shall see that the majority of users in our user study found the results by Social Pulse for these queries to be very informative. This suggests a potentially interesting use case of Social Pulse, where the user does not have a crystalized a-priori expectation of the results and the search engine returns a set of results that have been filtered socially.

### 4.4.4 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 repeated our analysis for top-5 and top-3 results, both for Google and Bing. The results for Bing exhibited the same trend as Google, so we focus on presenting the results for Google. Figures 4-5 and Tables 7-8 show the results. Overall we observe that our results are consistent, in terms of showing small overlap between Google and Social Pulse. Furthermore, when restricting the analysis to fewer than 10 results, the overlap decreases.

We also carried out another experiment in which we took the bottom five results from the top-6 results produced by Social Pulse and treated them as if they were the top-5 results of Social Pulse. We then compared these results to Google’s top-5 results. Through this experiment, we wanted to get a handle on the robustness of our conclusions to the variations in Social Pulse’s ranking function and the errors in tweet selection. We again found that the trends were preserved. We omit showing actual data.

### 4.4.5 Consistency with the TRM method

Recall our overview of the TRM method [53], given in Section 2. In order to study the consistency between our results with what one would obtain using the TRM method, we conducted another sen-

| Table 5: Queries exhibiting highest predictability. |
|---|---|---|
| **TRENDS** | **Social Pulse** | **Google** |
| Trends | Honda | Antibiotics |
| Manual | coup | education |

| Table 6: Queries exhibiting lowest predictability. |
|---|---|---|
| **TRENDS** | **Social Pulse** | **Google** |
| Trends | coup | iPhone |
| Manual | education | globalization |

| Table 7: AUC results for CROSSLEARNCOMPARE comparing Google and Social Pulse for top-5 results. |
|---|---|---|---|---|---|
| **Google to Social Pulse** | **Social Pulse to Google** |
| **TRENDS** | 0.9606 | 0.9035 |
| **Manual** | 0.9394 | 0.9647 |

| Figure 4: Comparing Google to Social Pulse for top-5 results |
|---|---|---|
| **(a) TENSORCOMPARE on (TRM)** | **(b) TENSORCOMPARE on (MANUAL)** |

| Figure 5: Comparing Google to Social Pulse for top-3 results |
|---|---|---|
| **(a) TENSORCOMPARE on (TRM)** | **(b) TENSORCOMPARE on (MANUAL)** |

| Table 8: AUC results for CROSSLEARNCOMPARE comparing Google and Social Pulse for top-3 results. |
|---|---|---|---|---|---|
| **Google to Social Pulse** | **Social Pulse to Google** |
| **TRENDS** | 0.9644 | 0.9986 |
| **Manual** | 0.5497 | 0.6930 |

| Table 9: Similarity from centroids (TRENDS) |
|---|---|---|---|
| From Google | To Google centroid | To Social Pulse centroid |
| From Social Pulse | 0.0143 | 0.1047 |
| From Social Pulse | 0.0464 | 0.1005 |
a well justified reason why they selected a particular answer, minimizing random responses and other forms of noise.

Considering budget for the study, a subset of the queries were used. Both TRENDS and MANUAL queries were included; the reader can see the complete list in Fig. 7. A HIT was created for every query and each of the top-10 search results for the query. We asked every HIT to be judged by ten users.

5.2 Inter-User Agreement

To ensure there is consistency in the judgments provided by the users, we measured the inter-user agreement using the *Fleiss’ kappa* ($\kappa$) value.

Table 11: Inter-user agreement. A column of this table provides the number of search results that were judged exactly by the corresponding number of users as well as the $\kappa$ value.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Google</th>
<th>Social Pulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kincade</td>
<td>11.3</td>
<td>7.1</td>
</tr>
<tr>
<td>ARI</td>
<td>13.5</td>
<td>9.5</td>
</tr>
<tr>
<td>Coleman-Liau</td>
<td>12.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Flesch Index</td>
<td>52.9/100</td>
<td>71.2/100</td>
</tr>
<tr>
<td>Fog Index</td>
<td>14.2</td>
<td>10.0</td>
</tr>
<tr>
<td>SMOG-Grading</td>
<td>12.5</td>
<td>9.6</td>
</tr>
<tr>
<td>Lix (school year 9)</td>
<td>49.3</td>
<td>42.2</td>
</tr>
<tr>
<td>Lix (school year 7)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Readability of results.

5.3 Sanity Checks

To further increase our confidence in the conclusions we arrive at, we did two sanity checks: i) visual inspection of the tweets in the result sets, and ii) their quantitative evaluation. We give below the results of both.

5.3.1 Visual Inspection

We examined top tweets for which there was high agreement amongst the judges as well as those tweets that had split judgments. Figures 7-9 show the top tweets from the two categories for which we had eight judgments each. It is readily apparent from these figures that the users were quite diligent in arriving at their decisions.
character, and we concatenate the snippets of each search engine style of the same set of results that we use for the user study. Due to the nature of the literature and applied in practice [20].

putes seven metrics that have been extensively discussed in the literature. To quantitatively answer this question, we put the tweets in our result sets containing a lot of misspellings and illegible terms. But does this belief hold water when we focus on highly retweeted tweets? To quantitatively answer this question, we put the tweets in our result sets through the unix style tool. Given a piece of text, this tool computes seven metrics that have been extensively discussed in the literature and applied in practice [20].

We conducted this study for Google and Social Pulse, for the same set of results that we use for the user study. Due to the nature of the style tool, we strip the snippets off any non alpha-numeric character, and we concatenate the snippets of each search engine into a longer passage, and apply style to it. The results are shown in Table 12.

It is not surprising that tweets score lower than Web snippets. The latter are derived from Web pages that are generally written more formally whereas communication on Twitter is relatively informal. Note also that a lower value of a readability metric does not automatically imply lower understandability of the content. For example, the most popular novels are written at the 7th-grade level and people read for recreation texts that are two grades below their actual reading level [30]. Interestingly, we see from Table 12 that Lix pegs the readability of the result tweets at the 7th-grade level.

5.4 Conclusions from the User Study

We can now finally present the results of our user study. Figure 10 summarizes them. We have plotted the usefulness index separately for each of the queries. For computing the usefulness index for a query, we consider every search result for a query for which we could get at least four judgments. We then check if a strict majority of users have judged the result to be informative for the given query. Note that “hard to tell” is treated as “not informative” for this purpose. The majority votes are then averaged over distinct search results for a specific query. Since the inter-user agreement is quite good according to Fleiss’ kappa, the majority vote is a good indicator of the result quality.

Overall, Fig. 10 demonstrates that most of the users found a large portion of Social Pulse’s results informative with respect to the query in question, regardless of the query category (TRENDS or MANUAL). This finding is remarkable given the fact that the sole signal we use in order to discover and rank these results is the number of retweets.

6. CONTRIBUTIONS AND FUTURE WORK
Our major contributions in this work are as follows:

1. Through a rigorous analysis of real data from Google, Bing, and Twitter, we showed that a search engine built using even simple social signals like retweet count can surface tweets whose content is quite different from those provided by the current search engines to the Web users. Our extensive user study demonstrated that not only is this content different, but can also be very informative.

These findings have direct, practical ramifications. Given the central role the commercial search engines play in arbitrating what information is seen by the citizens and the importance of ready access to diverse view points for inculcating an informed citizenry, it behooves the commercial search engines to conduct studies similar to ours in their own infrastructure. They certainly have the financial and computing resources as well as ready availability of data for conducting such explorations and provide the choice of access to diversity to the citizens.

2. By successfully reusing the methodology and tools, introduced in [2], for carrying out the present investigation of distinctiveness of social network content from the Web content, we reinforced the power of data mining to be able to abstract meaningful insights from massive amount of data.

3. We generated data sets that other researchers might be able to use for making their own discoveries.

Looking ahead, the Web search engines can start providing search results from social networks in two phases (assuming they see sufficient business imperative for it):

1. Add a social tab to their search result page. Research can contribute by refining algorithms for global and personalized ranking of social results as well as addressing the related infrastructure and environmental issues such as trust and privacy. Other topics for fruitful research include drill down into the differentiating attributes of social results and characterizing the phenomena that underlie the differences [53].

2. Intermix the social results with the Web results. Research can contribute by building comprehensive diversity models as well as evaluating and extending algorithms for diversifying search results [15][17].

7. REFERENCES


