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 nurtures an informed citizenry. Google and Bing have emerged as the duopoly that largely arbitrates which 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 gain different perspectives by using 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, our user study found that these results were quite informative. 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.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; H.3.3 [**Information Search and Retrieval**]: Search process

Keywords

Web search; social media search; search engine; search result comparison; Google; Bing; Twitter

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 [19]. The World Wide Web is now widely recognized as the universal information source. Content representing diverse perspectives exist on the Web, on almost on any topic. However, this does not automatically ensure that citizens encounter them [46].

Search engines have become the primary tool used to access the web content [38]. 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).

KDD '15 Sydney, NSW, Australia

DOI: http://dx.doi.org/10.1145/2783258.2788571.

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 most of the clicks as users rarely look at results at lower positions [18, 24]. 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. The implication is that citizens may no longer be able to gain 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 [33]. 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. Users have judged those results to be quite informative.

We used Twitter in our study because it is still possible to selectively crawl Twitter.

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 gives the experimental setup and Section 5 presents the results of our analysis using data from Google, Bing, and Twitter. Section 6 presents the user study for assessing the usefulness of our findings. We conclude with a summary and future directions in Section 7.

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 [49].

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 [14]. Around the same time, Selberg and Etzioni found that each of Galaxy, Infoseek, Lycos, OpenText, Webcrawler and Yahoo returned mostly unique results [43]. 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 [21]. 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% [8]. 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 [30]. 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% [44].

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 [52] 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 [36]. 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 [2]. 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 the access to diverse perspectives becomes harder.

Contrary to the rich literature on overlap between the web search engines results, the only prior work we could find on overlap between web and social search results appears in Section 5 of [49] (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) [9]. 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 [4, 25]. 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 5.

2.2 Social Search

In addition to being considered a social media and a social network [28], Twitter may also be viewed as a 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 [11]. 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".[1]

The prevailing perception is that much of the content found on Twitter is of low quality [3] and the keyword search as provided by Twitter is not effective [48]. 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 [12, 16, 48, 51]. 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 [50]. But, of course, one needs to remove socware and other spam before using retweet count [35, 39, 42] Other approaches include using the presence of a URL as an indicator [3], link analysis on the follows and retweet graphs [40, 53], clustering taking into account the size and popularity of a tweet, its audience size, and recency [29], and the semantic approaches including topic modeling [54]. See overviews in [51, 54] for additional references.

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

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 10:42AM, 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,

[1]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.

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. Google is also said have struck a deal with Twitter that will allow tweets to be shown in Google search results sometime during this year.

There is also research on how web search can be improved using signals from Twitter. For example, Rowlands et al. [41] 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 tweeterer may serve to weight both annotations and query-independent popularity. Similarly, Dong et al. [15] 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 $\underline{\mathbf{X}}$, where (query, result, time, search engine) are the four modes [27]. 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 PARAFAC decomposition [23] 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 $\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r, \mathbf{d}_r$ have been normalized with their scaling absorbed in λ_r. For compactness, the decomposition is represented as matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$. The decomposition of $\underline{\mathbf{X}}$ to $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ gives a low rank embedding of queries, results, timings, and search engines respectively. The factor matrix $\mathbf{D}$ 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 $\mathbf{D}$ 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 between 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, multiple points are shown along the diagonal for the sake of generality.

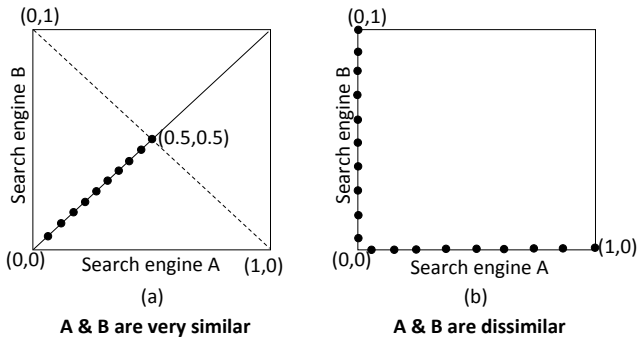


Figure 1: Visualization guide for TENSORCOMPARE.

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. 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 $\mathbf{D}(1, j)x + \mathbf{D}(2, j)y = 1$, 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).

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. EXPERIMENTAL SETUP

We next describe the experimental setup of the empirical study we performed, applying the tools just described.

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. [7, 37, 45, 47]). 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 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 [10]. Many of them are named entities, which constitute a significant portion of what people search. The total number of queries was limited by the budget available for the study.

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

We probed the search engines during June-July 2014 with the same set of queries at the same time of the day for 21 (17) days for the TRENDS (MANUAL) set. 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.

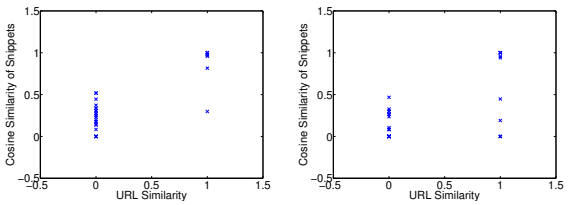
In all cases, we recorded the top-k results. The value of k is set to 10 by default, except in the experiments studying the sensitivity of results to the value of k. Every time, we ran the same code from the same machine having the same IP address to minimize noise in the results. Because we were getting 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 [22].

4.3 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 [34]. The alternative of using URL representation must first address the well-known problems arising from short URLs [5], un-nomalized URLs [31, 32], and different URLs with similar text [6]. 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 [34]. In the case of Social Pulse, the entire text of a tweet (including hashtags and URLs, if any) 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 [49].

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; binary features did not change the trends. For CROSSLEARNCOMPARE, we keep the top-n words and have binary features. Finally, we note that the distribution of the snippet lengths for Google, Bing, and Social Pulse was almost identical for all the queries we tested. This ensures a fair comparison between them.

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 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

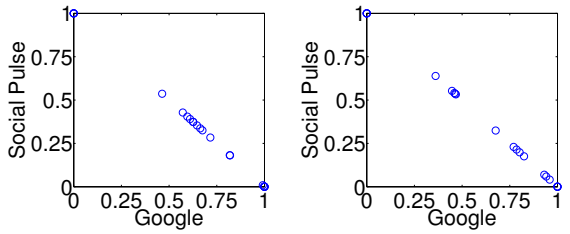


(a) TRENDS query set (b) MANUAL query set

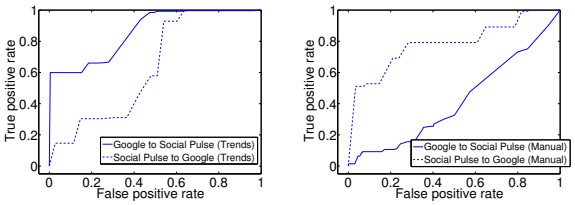
Figure 2: Comparing URL similarity with snippet similarity

We see that for most of the queries for which the snippet similarity is low, the results point to different documents. On the other hand, when this similarity is high, the documents are identical. In both TRENDS and MANUAL, there exist some outliers with pointers to identical documents yet dissimilar snippets. Yet, overall, Fig. 2 indicates that snippets are good vehicles 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).



(a) TENSORCOMPARE for TRENDS (b) TENSORCOMPARE for MANUAL

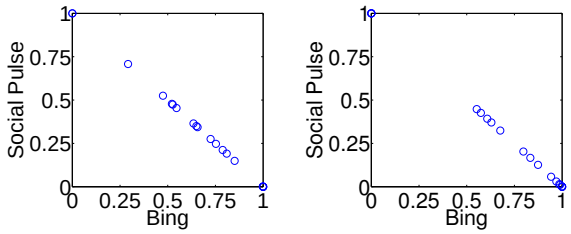


(c) CROSSLEARNCOMPARE for TRENDS (d) CROSSLEARNCOMPARE for MANUAL

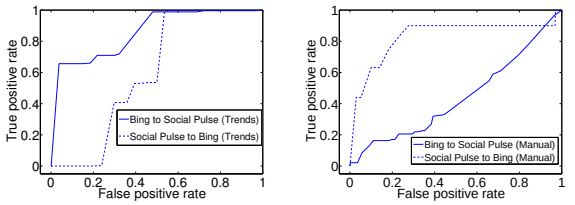
Figure 3: Social Pulse vs. Google for top-10 results

	TRENDS →	TRENDS ←	MANUAL →	MANUAL ←
Google- Social Pulse	0.86	0.64	0.42	0.78

Table 3: AUC for CROSSLEARNCOMPARE comparing Google and Social Pulse for top-10 results.



(a) TENSORCOMPARE for TRENDS (b) TENSORCOMPARE for MANUAL



(c) CROSSLEARNCOMPARE for TRENDS (d) CROSSLEARNCOMPARE for MANUAL

Figure 4: Social Pulse vs. Bing for top-10 results

	TRENDS →	TRENDS ←	MANUAL →	MANUAL ←
Bing- Social Pulse	0.86	0.60	0.44	0.83

Table 4: AUC for CROSSLEARNCOMPARE comparing Bing and Social Pulse for top-10 results.

5. FINDINGS

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

5.1 Social Pulse Versus Google

Figure 3 and Table 3 show the results. We see in Figs. 3(a), 3(b):

1. There exist 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. 3(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 3(c) and AUC 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 users might not retweet much the readily-available, main-stream content on popular topics.

In contrast, for the MANUAL dataset in Fig. 3(b), 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.

5.2 Social Pulse Versus Bing

We repeated the preceding analysis, but by using Bing search results rather than Google this time. Figure 4 and Table 4 show the results. These results are qualitatively similar to those obtained using Google search results, which is not surprising given the earlier finding 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.

5.3 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.

	Google → Social Pulse	Social Pulse → Google
TRENDS	SpongeBob SquarePants Albert Einstein Tottenham Hotspur F.C.	Oprah Winfrey Maya Angelou Albert Einstein
MANUAL	self-driving car gay marriage San Francisco	World cup gay marriage World bank

Table 5: Queries exhibiting highest predictability.

	Google → Social Pulse	Social Pulse → Google
TRENDS	Honda Antibiotics Frozen	Game of Thrones Skrillex Martini
MANUAL	coup education globalization	coup iPhone poverty

Table 6: Queries exhibiting lowest predictability.

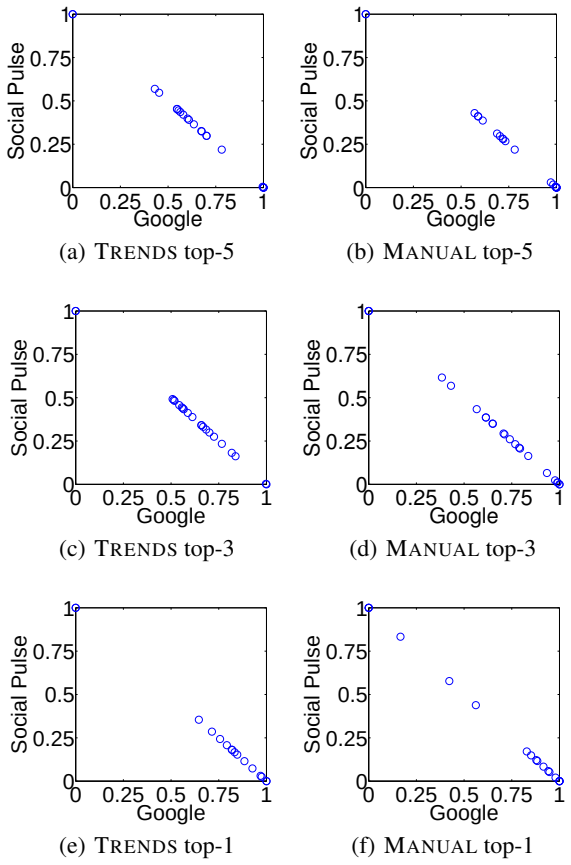
We see that the timely queries, like *World cup* or *gay marriage*, have high mutual predictability. Indeed, timeliness creates relevance; the same information gets retweeted and clicked a lot. Queries like *Maya Angelou* and *Albert Einstein* are also highly mutually predictable, in part because people 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.

5.4 Sensitivity Analysis

We repeated our analysis for top-5, top-3 and top-1 search results. The results for Bing exhibited the same trend as Google, so we focus on presenting the results for Google. Figures 5 and Table 7 show the results. Overall we observe that our results are consistent, in terms of showing small overlap between Google and Social Pulse.

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



(a) TRENDS top-5 (b) MANUAL top-5

(c) TRENDS top-3 (d) MANUAL top-3

(e) TRENDS top-1 (f) MANUAL top-1

Figure 5: TENSORCOMPARE sensitivity

Google- Social Pulse	TRENDS →	TRENDS ←	MANUAL →	MANUAL ←
top-10	0.86	0.64	0.42	0.78
top-5	0.87	0.70	0.39	0.66
top-3	0.86	0.50	0.35	0.69
top-1	0.79	0.98	0.50	0.53

Table 7: AUC for CROSSLEARNCOMPARE comparing Google and Social Pulse for different number of top results.

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.

5.5 Consistency With TRM Method

Recall our overview of the TRM method [49], 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 sensitivity experiment. We first apply tensor analysis to the Google and Social Pulse results to obtain their condensed representations. We then compute the centroids for the Google and the Social Pulse results topics, and for every result from Google and Social Pulse (for all queries and days), we compute its cosine similarity to each centroid. While calculating the centroids, we ignore topics that are shared between Google and Social Pulse and keep those that lie on the $(0, 1)$ and $(1, 0)$ points of the TENSORCOMPARE plots. We present the results of this experiments in Table 8.

		To Google centroid	To Social Pulse centroid
TRENDS	From Google result	0.20	0.10
	From Social Pulse result	0.05	0.10
		To Google centroid	To Social Pulse centroid
MANUAL	From Google result	0.22	0.10
	From Social Pulse result	0.05	0.11

Table 8: Similarity from centroids

We again see that Google results in both query sets are more similar to the Google centroid, and Social Pulse results to the Social Pulse centroid. This analysis, this time employing a different method, further reinforces the conclusion that the social search results can be quite different from the conventional Web search results.

6. USER STUDY

So far, we have discovered that the results of Social Pulse are different from Google and Bing. However, one might wonder whether these different results are actually useful, particularly given the apprehension that the content found on Twitter is of low quality [3]. To that end, we conducted a user study on the Amazon Mechanical Turk platform, following the best practices recommended in [1]

6.1 HIT Design

Taking cue from the relevance judgment literature [13], the HIT (Human Intelligence Task) presented to the users consist of a query and a text representing a search result. The users are asked to select whether 1) the text is not informative, 2) the text is informative, or 3) it is hard to tell. They are then asked to explain their answer; any HIT that did not provide this explanation was rejected. Figure 6 shows a sample HIT.

We used the phrase "informative" rather than "relevant" in the instructions, after some initial testing. The choice "not informative" was placed above the positive one to avoid biasing the user's response towards the positive answer. Requiring users to explain their answer turned out to be important: users were forced to have a well justified reason why they selected a particular answer, minimizing random responses and other forms of noise.

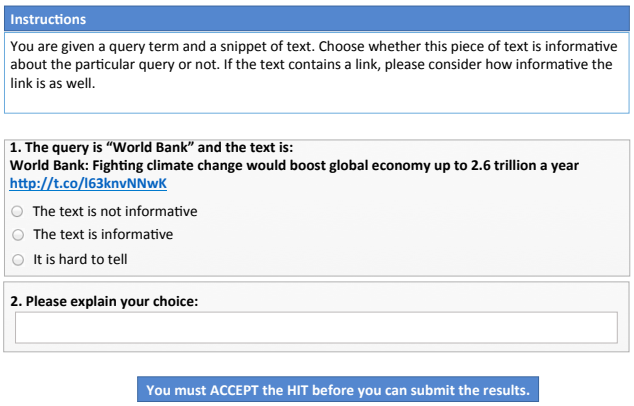


Figure 6: A sample HIT

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. 10. 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.

6.2 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* test [20]. In a nutshell, Fleiss' kappa (κ) is a number that indicates the degree of agreement between judges that is statistically significant and not attainable by chance. Its maximum value (for perfect agreement) is 1, and where there is no agreement, it can also take negative values. In exploratory tasks such as ours, a value in the range of 0.2-0.4 shows reasonable agreement and confidence on the results.

Although we had sought 10 judgments for every HIT, the actual deployment yielded the number of good judgments ranging from 4-10. Table 9 shows the κ values for our user study. A column of this table shows the number of search results that were judged exactly by the corresponding number of users as well the κ value. Thus, the column 1 of this table indicates that for twelve of the search results each was judged by exactly four users. We observe that κ is reasonably good in all cases, signifying good inter-user agreement.

# Judges	4	5	6	7	8	9	10
# Results	12	70	92	91	51	25	5
κ	0.19	0.29	0.27	0.26	0.32	0.43	0.22

Table 9: 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 κ value.

6.3 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.

6.3.1 *Visual Inspection*

We examined top tweets for which there was high agreement amongst the judges as well as those tweets that had split judgments.

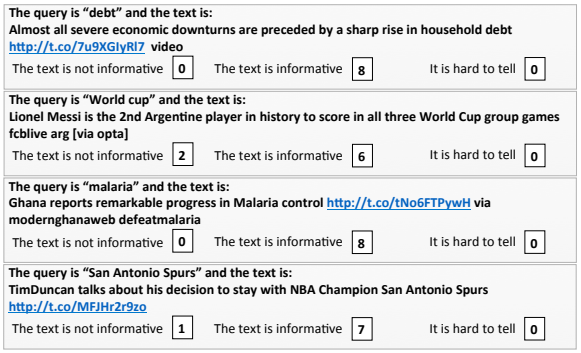


Figure 7: Informative results with high judge agreement

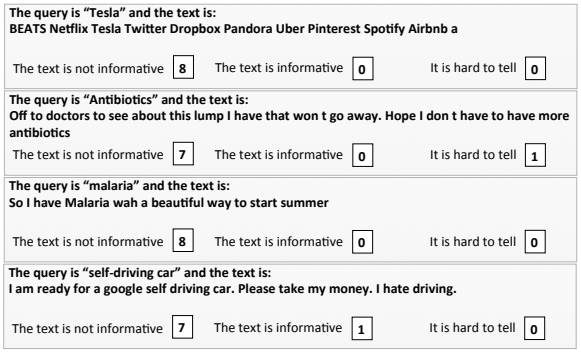


Figure 8: Not informative results with high judge agreement

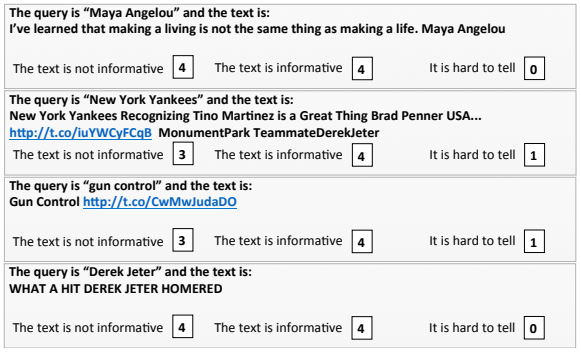


Figure 9: Results with poor judge agreement

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.

6.3.2 *Readability of the Result Tweets*

It is a common belief that tweets are usually of bad quality, 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 [17].

Metric	Google	Social Pulse
Kincaid	11.3	7.1
ARI	13.5	9.5
Coleman-Liau	12.9	11.4
Flesch Index	52.9/100	71.2/100
Fog Index	14.2	10.0
SMOG-Grading	12.5	9.6
Lix	49.3 (school year 9)	42.2 (school year 7)

Table 10: Readability of results.

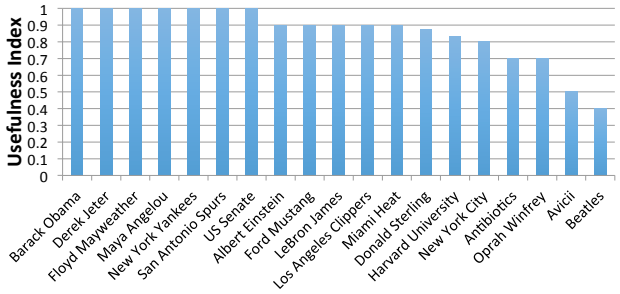
We conducted this study for Google and Social Pulse, for the same set of results that we use for the user study. Due to the nature of the `style` tool, we strip the snippets off any non alpha-numeric character, and we concatenate the snippets of each search engine into a longer passage, and apply `style` to it. The results are shown in Table 10.

It is not surprising that tweets score lower than Web snippets. The latter are derived from Web pages that are generally written much 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 [26]. Interestingly, we see from Table 10 that Lix pegs the readability of the result tweets at the 7th-grade level.

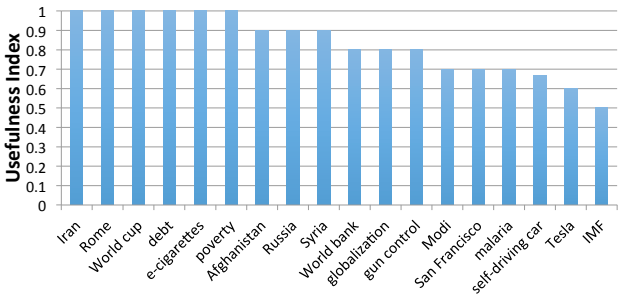
6.4 Results of the User Study

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.



(a) TRENDS



(b) MANUAL

Figure 10: Usefulness index of search results produced by Social Pulse for various queries

7. SUMMARY 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 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, 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 [51].

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 [13, 33].

Acknowledgements This work was partially done at the erstwhile Microsoft Research in Silicon Valley. E. Papalexakis was partially supported by National Science Foundation Grant IIS-1247489.

8. REFERENCES

[1] *Amazon Mechanical Turk, Requester Best Practices Guide*. Amazon Web Services, June 2011.

[2] R. Agrawal, B. Golshan, and E. Papalexakis. A study of distinctiveness in web results of two search engines. In *24th international conference on World Wide Web, Web Science Track*. ACM, 2015.

[3] O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar. Detecting uninteresting content in text streams. In *SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[5] D. Antoniades, 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*, pages 715–724. ACM, 2011.

[6] Z. Bar-Yossef, I. Keidar, and U. Schonfeld. Do not crawl in the DUST: different urls with similar text. *ACM Transactions on the Web*, 3(1):3, 2009.

[7] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee. Precise tweet classification and sentiment analysis. In *IEEE/ACIS 12th international conference on Computer and Information Science*, pages 461–466. IEEE, 2013.

[8] K. Bharat 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, 1998.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[10] A. Broder. A taxonomy of web search. *ACM Sigir forum*, 36(2):3–10, 2002.

[11] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin. Earlybird: Real-time search at Twitter. In *IEEE 28th international conference on Data Engineering*, pages 1360–1369. IEEE, 2012.

[12] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *20th international conference on World Wide Web*, pages 675–684. ACM, 2011.

[13] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 web track. Technical report, NIST, 2011.

[14] W. Ding and G. Marchionini. A comparative study of web search service performance. In *ASIS Annual Meeting*, volume 33, pages 136–42. ERIC, 1996.

[15] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *19th international conference on World Wide Web*, pages 331–340. ACM, 2010.

[16] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *23rd international conference on Computational Linguistics*, pages 295–303. Association for Computational Linguistics, 2010.

[17] W. DuBay. *The principles of readability*. Impact Information, 2004.

[18] E. Enge, S. Spencer, J. Stricchiola, and R. Fishkin. *The art of SEO*. O'Reilly, 2012.

[19] Federal Communications Commission. Editorializing by broadcast licensees. Washington, DC: GPO, 1949.

[20] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[21] S. Gauch and G. Wang. Information fusion with profusion. In *1st World Conference of the Web Society*, 1996.

[22] A. Hannak, 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*, pages 527–538. ACM, 2013.

[23] R. A. Harshman. Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis. Technical report, UCLA, 1970.

[24] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 133–142. ACM, 2002.

[25] U. Kang, 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*, pages 316–324. ACM, 2012.

[26] G. R. Klare and B. Buck. *Know Your Reader: The scientific approach to readability*. Heritage House, 1954.

[27] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[28] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *19th international conference on World Wide Web*, pages 591–600. ACM, 2010.

[29] H. W. Lauw, A. Ntoulas, and K. Kenthapadi. Estimating the quality of postings in the real-time web. In *Proc. of SSM conference*, 2010.

[30] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.

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

[32] T. Lei, 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*, pages 611–620. ACM, 2010.

[33] V. Maltese, F. Giunchiglia, K. Denecke, P. Lewis, C. Wallner, A. Baldry, and D. Madalli. *On the interdisciplinary foundations of diversity*. University of Trento, 2009.

[34] M.-C. Marcos 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, 2010.

[35] J. Martinez-Romo and L. Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000, 2013.

[36] W. Meng, C. Yu, and K.-L. Liu. Building efficient and effective metasearch engines. *ACM Computing Surveys*, 34(1):48–89, 2002.

[37] K. Nishida, T. Hoshide, and K. Fujimura. Improving tweet stream classification by detecting changes in word probability. In *35th international ACM SIGIR conference on Research and development in information retrieval)*, pages 971–980. ACM, 2012.

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

[39] M. S. Rahman, T.-K. Huang, H. V. Madhyastha, and M. Faloutsos. Efficient and scalable socware detection in online social networks. In *USENIX Security Symposium*, pages 663–678, 2012.

[40] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *Machine learning and knowledge discovery in databases*, pages 18–33. Springer, 2011.

[41] T. Rowlands, D. Hawking, and R. Sankaranarayana. New-web search with microblog annotations. In *19th international conference on World Wide Web*, pages 1293–1296. ACM, 2010.

[42] I. Santos, I. Miñambres-Marcos, C. Laorden, P. Galán-García, A. Santamaría-Ibirika, and P. G. Bringas. Twitter content-based spam filtering. In *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, pages 449–458. Springer, 2014.

[43] E. Selberg and O. Etzioni. Multi-service search and comparison using the metacrawler. In *4th international conference on World Wide Web*, 1995.

[44] A. Spink, 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, 2006.

[45] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM, 2010.

[46] N. J. Stroud and A. Muddiman. Exposure to news and diverse views in the internet age. *ISJLP*, 8:605, 2012.

[47] H. Takemura and K. Tajima. Tweet classification based on their lifetime duration. In *21st ACM international conference on Information and knowledge management*, pages 2367–2370. ACM, 2012.

[48] K. Tao, F. Abel, C. Hauff, and G.-J. Houben. Twinder: a search engine for twitter streams. In *Web Engineering*, pages 153–168. Springer, 2012.

[49] J. Teevan, 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*, pages 35–44. ACM, 2011.

[50] I. Uysal and W. B. Croft. User oriented tweet ranking: a filtering approach to microblogs. In *20th ACM international conference on Information and knowledge management*, pages 2261–2264. ACM, 2011.

[51] W. M. Webberley. *Inferring Interestingness in Online Social Networks*. PhD thesis, Cardiff University, 2014.

[52] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *18th ACM conference on Information and knowledge management*, pages 87–96. ACM, 2009.

[53] M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim. Finding interesting posts in twitter based on retweet graph analysis. In *35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1073–1074. ACM, 2012.

[54] M.-C. Yang and H.-C. Rim. Identifying interesting twitter contents using topical analysis. *Expert Systems with Applications*, 41(9):4330–4336, 2014.