The World Conversation: Web Page Metadata Generation From Social Sources

Omar Alonso
Microsoft
omalonso@microsoft.com

Sushma Bannur*
Facebook
bnsushma@fb.com

Kartikay Khandelwal
Microsoft
kartikk@microsoft.com

Shankar Kalyanaraman*
Facebook
kshankar@fb.com

ABSTRACT

Over the past couple of years, social networks such as Twitter and Facebook have become the primary source for consuming information on the Internet. One of the main differentiators of this content from traditional information sources available on the Web is the fact that these social networks surface individuals' perspectives. When social media users post and share updates with friends and followers, some of those short fragments of text contain a link and a personal comment about the web page, image or video. We are interested in mining the text around those links for a better understanding of what people are saying about the object they are referring to. Capturing the salient keywords from the crowd is rich metadata that we can use to augment a web page. This metadata can be used for many applications like ranking signals, query augmentation, indexing, and for organizing and categorizing content. In this paper, we present a technique called social signatures that given a link to a web page, pulls the most important keywords from the social chatter around it. That is, a high level representation of the web page from a social media perspective. Our findings indicate that the content of social signatures differs compared to those from a web page and therefore provides new insights. This difference is more prominent as the number of link shares increase. To showcase our work, we present the results of processing a dataset that contains around 1 Billion unique URLs shared in Twitter and Facebook over a two month period. We also provide data points that shed some light on the dynamics of content sharing in social media.

Categories and Subject Descriptors

H:0 [**Data**]: General; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

*Work done while author was affiliated with Microsoft.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2745397.

Keywords

Social media, Twitter, Facebook, metadata, web page augmentation, annotation.

1. INTRODUCTION

The remarkable rise in the use of social networks such as Twitter and Facebook has become a significant driver of Internet traffic towards websites and web pages. While using social media, users not only share content and links to web pages on their social network, but they also provide additional information about the nature of those links. For instance, a user tweeting a link to a movie page is likely to add annotations that provide some qualitative signal concerning the movie. Questions along the lines of "Why is this entity trending?" or "What are people saying about this web page?" are often difficult to answer because of the large amounts of noise present in the data. Further, for the links corresponding to web pages with scarce textual content, such as videos and images, social text surrounding these links can provide valuable understanding.

As a motivating scenario, say that the following video about the San Francisco Airport (SFO) accident is shared across Twitter and Facebook. Searching for the airport's name on a search engine shows the traditional snippet and the article's title but it won't capture what the *crowd* is conversing on different social networks. We are not interested in the sentiment of the conversation but rather the most salient terms used regardless of the point of view, like the example on Figure 1. How can we extract the most pertinent keywords from all the comments in Twitter and Facebook about this specific web page? One way would be to search on Twitter but we often get other results and if the event is popular, there will be thousands of similar tweets. Facebook would give us all the posts restricted to our own subgraph.

Users' comments or *social annotations* are human generated content and mining them can provide powerful insights about the link, especially with regard to dynamic content being shared (e.g., viral videos, hot deals, breaking news, etc.). Besides world events like disasters, we show another example of how significant comments can be in Figure 2. Here we show the social signatures presented as a tag cloud on the announcement that Ben Affleck was going to play Batman based on a heavily shared link. In the figure the size of the terms corresponds to the strength of the term.

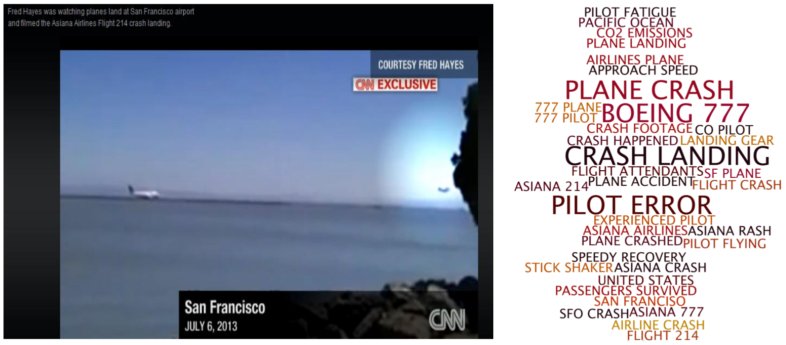


Figure 1: SFO-Asiana event. The image on the left shows the video caption from the CNN link. On the right, a Wordle visualization of the social signatures extracted from social networks.
www.cnn.com/videos/us/2013/07/07/vo-plane-sf-plane-crash-on-cam.courtesy-fred-hayes

As a specific application to web search, social annotations can provide not just additional context but also offer snapshots in time by capturing the vocabulary of these social conversations. In web search, mining anchor text has been an area of active research and product development. More recently, there has been work on utilizing social annotations for improving web search and understanding web page metadata (social annotations, anchor text and social queries) for related applications. In the context of social media, we propose the notion of a *social anchor text* as a short and concise summary provided by a user about a web page. In the case of Twitter, the social anchor text is what is left on the tweet after extracting the link, profile handles and hashtags. For Facebook, this corresponds to the posts and comments associated with a link.

We define a *social signature* as the set of tokens that provide a high level representation of the web page using social data. This is similar to the notion of *lexical signatures* [17] but using the social anchor text instead of the content of a web page.

In this paper we present the computation of social signatures, condensing the most important keywords for a shared link. The goal of this computation is to obtain a list of signatures that can give insight into the social buzz around the link and can help answer questions similar to the ones presented earlier. This new generated metadata can later be used in many scenarios such as search results ranking, indexing, and content organization. Figure 3 presents a high level overview of our approach.

Our methods in this paper are simple and this is intentional. We want our method of deriving social signatures to be scalable at web latencies, therefore giving us little leeway in sophisticated algorithms that may only achieve marginally higher relevance while losing out on efficiency.

We evaluate social signatures in two ways. First, we compare the social signatures with the web page text for the shared links and present the new insights gained from social comments. Second, we use human judgments to assess the quality of the generated signatures given a web page.

The contributions of this paper are:

1. Design and implementation of an efficient system for computing social signatures for web links based on social media at scale.
2. A comparison of social signatures with web page content.
3. A data analysis on links shared on Twitter and Facebook.

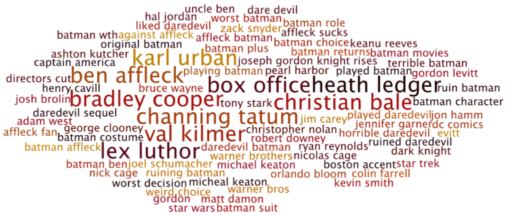


Figure 2: Wordle based on social signatures about the new Batman movie. It is very interesting to see all the potential candidates that are mentioned for the role as well as characters that are not related at all (e.g., Lex Luthor).
movies.yahoo.com/news/ben-affleck-set-play-batman-man-steel-sequel-011253917.html

This paper is organized as follows. In the next section we present an overview of the related work in this area. We then describe the dataset characteristics used to demonstrate our system. The Methodology section explains the methods and techniques used for computing social signatures. The Evaluation section illustrates the experimental setup and presents our findings, new insights gained, comparison with web page content and human evaluation. We describe potential applications based on social signatures and finalize with conclusions and future work.

2. RELATED WORK

In this section, we discuss previous research as it relates to our work in the context of social anchor text and implementation methodology. Anchor text used to describe links

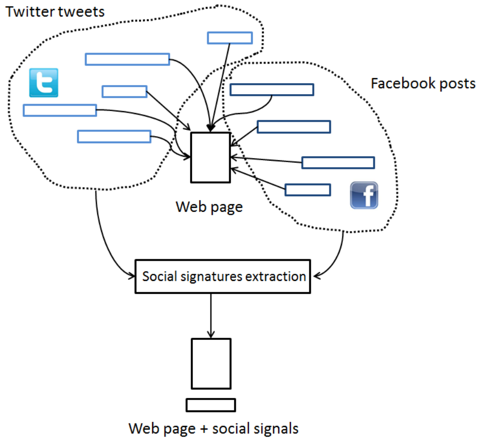


Figure 3: Web page metadata augmentation by mining tweets from Twitter and posts from Facebook as social signatures.

or URLs in web pages has been studied as a useful summarization primitive to improve search quality. Eiron and McCurley [6] showed that anchor text, in addition to titles, resemble typical user queries and are an important feature for ranking and relevance. Anick [3] has looked at how anchor text surrounding an incoming hyperlink can be re-purposed as a lexical resource used for NER (Named-Entity Recognition) on the target page, as well as for improving query intent and precision. Amitay *et al.* [2] looked at the different keywords users input during query sessions and generate anchor representations to enhance document-based features used to measure search quality. Zhou *et al.* [21] incorporate user-browsing activity into building anchor document representations to improve the ranking signal anchor text can provide for search. Wu *et al.* [20] have compared various techniques using anchor text in search and retrieval over baseline approaches.

In terms of the methodology for implementing annotations (user-generated or from anchor text) for search results, researchers and practitioners have investigated contrasting techniques. While Fujii [8] modeled anchor text by treating each snippet individually, Metzler *et al.* [12] aggregated anchor text to create a single piece of annotation. Noll and Meinel [15] were among the first to evaluate social annotations for enhancing search results. They compare three forms of metadata social annotations, anchor text and query keywords. They find in their evaluations that social annotations, in the form of tags provided by del.icio.us users, tend to perform better than anchor text on all three counts. Lee and Croft [11] analyzed the effect of query-dependent and query-independent features extracted from a small set of URLs shared on social domains, and their social anchor document representations. Boydell and Smyth [4] consider community-specific search engines where users' search terms are used to personalize the snippets appearing in the search results. However, the snippets in their work are derived from within the document itself and do not provide information different from the web page.

In work that is perhaps closest in intent to ours, Ferragina and Scaiella [7] propose improving the quality of snippets in search results by augmenting them with related Wikipedia entities mentioned in the text. We envision their work as being orthogonal to and compatible with our approach – using socially generated snippets to add freshness and pertinence to search results. We cast similarly the work of Raux *et al.* [18] who use the text around shared URLs in tweets to identify trending topics.

Mishne and Lin [13] hinted at the potential value of social anchor text in their preliminary study based on tweets containing links to web pages. They show that social conversations contribute significantly by way of novelty to the overall anchor document representation, especially for frequently shared URLs. We develop this idea further and report on our experience in building a social annotation system on top of existing search infrastructure.

In previous work social annotations have come to mean some form of signal from a user's social circle indicating a personal recommendation/approval to a search result. Muralidharan *et al.* [14] looked at social annotations on Google's search engine results page. Specifically they evaluated how including any additional context situating a search result influenced the users' ability to process results from web searches. Similarly, Pantel *et al.* [16] also looked at the utility of adding such snippets as judged by user studies. They found that the utility of showing such annotations varies widely and depends specifically on the expertise and social distance of the annotators.

This is different from our intent in this work, which is to aggregate a social signature of text and metadata. By social signature here, we look beyond just the user's immediate social circle and think of it as metadata derived from the society at large. An entity-centric view on detecting the salience of entities within web documents is proposed by Gamon *et al.* [9].

A different perspective on link sharing, is the recent work by Gerlitz and Helmond [10] on the Like economy. Their work has a Facebook-only focus in the context of social buttons present in web pages.

3. DATASET CHARACTERISTICS

To demonstrate the implementation of our techniques, we consider the links shared on the two largest social networks, Facebook and Twitter. The dataset consists of URLs shared on these networks in the two month period from July 1^{st} 2013 to 31^{st} August 2013. Specifically, the dataset comprises the URLs and the raw text associated with every share. The raw URL is processed to extract the target URL corresponding to the landing web page, expanding the links from URL shortening services like `bit.ly`. The raw text associated with a URL is defined as the social anchor text. A language detection classifier is used to filter shares with social anchor text in non-English language. Thus, the dataset comprises landing web page URLs and the corresponding raw social anchor text in English language.

If we look at the data from a Twitter perspective, the dataset contains 931M (million) unique URLs extracted from 2B (billion) English language tweets that contain at least one link. If we look at the data from a Facebook perspective, the dataset corresponds to 116M unique URLs extracted from links shared by Facebook users. The dataset consists of shared URLs and the associated posts, comments in English language. It is observed that significant number of URLs shared on Facebook contain internal links in the social network corresponding to Facebook pages, profiles, events, groups, posts, questions, etc. In this study, we exclude all the shares that correspond to links within the Facebook network. Figure 4 shows the log-log plot of the distribution of the number of domains versus the frequency of each domain for the datasets. Figure 5 shows the same plot for URLs. We observe that both plots follow a power law with a very small fraction of URLs and domains being shared a large number of times while the majority of them are shared only a few times (< 10).

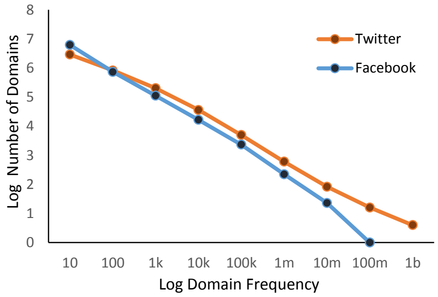


Figure 4: Log-log of number of domains vs. domain frequency.

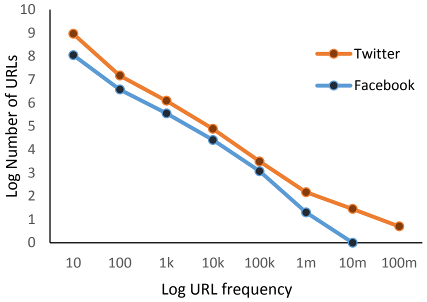


Figure 5: Log-log of number of URLs vs. URL frequency.

The top 5 domains shared on Twitter, correspond to: `twitter.com`, `instagram.com`, `facebook.com`, `ask.fm`, and `unfollowers.me` respectively. On the other hand the top 5 domains shared on Facebook correspond to: `youtube.com`, `ask.fm`, `instagram.com`, `buzzfeed.com` and `soundcloud.com`.

We also compare the social anchor text for a URL with the web page content for the URL. For this, we fetch the raw HTML content for a subset of the URLs in the Facebook and Twitter shared links dataset from a commercial search index. The subset comprises URLs with non-empty social signatures, explained in detail in Section 5. Web page content of 368K URLs is extracted corresponding to the Facebook share links subset. For Twitter, web page content of 1M URLs is extracted. Figure 6 and Figure 7 show the distribution of the domains and URLs for the subset of the shared link dataset with web-page content. The figures confirm that the subset follows the power law similar to the distribution in the original shared link datasets.

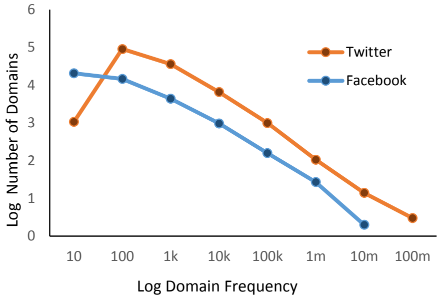


Figure 6: Log-log of number of domains vs. domain frequency for subset of URLs with web page content.

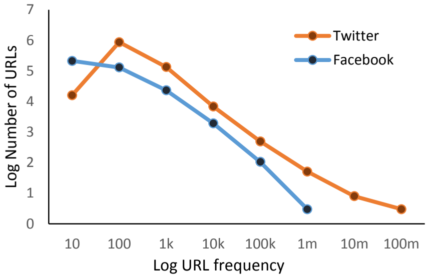


Figure 7: Log-log of number of URLs vs. URL frequency for subset of URLs with web page content.

4. COMPUTING SOCIAL SIGNATURES

In this section we discuss our technique for computing the social signature for a particular URL. We denote $U = \{u_1, \ldots, u_n\}$ as the set of all URLs seen in our data set and $d = \{d_1, \ldots, d_n\}$ the set of all documents where a document d_j represents the set of all social fragments $\{s_1, \ldots, s_m\}$ associated with URL u_j. A fragment for more than one URL can belong to multiple documents because multiple URLs can exist in a single social fragment. We denote $sig = \{sig_1, \ldots, sig_k\}$, as a list of N-grams and associated scores that constitute the social signature for URL u_j extracted using the text in document d_j. Using this notation, we can define our problem as aggregating all the social content related to a URL u_j to obtain document d_j, which is in turn used to generate a list of pairs sig_j. We limit the N-grams to $N = 2$.

4.1 Methodology

For each URL u_j, we generate the document d_j which aggregates all the social anchor text around that URL from different users. For Facebook, the social anchor text comprises of the text from posts and comments from users. For Twitter, the social anchor text comprises of the text in tweet, excluding the twitter user profile handles and hashtags mentioned in the tweet. As the text in the social fragments from users on social networks lacks structure as compared to typical web documents, eliminating noise and extracting meaningful content is one of the biggest challenges. We carefully choose specific rule based pattern matching and certain dictionary lookup based approaches to eliminate noise. The raw aggregated social anchor text is processed by eliminating stop words based on lookup on generic English language stop words and a social network specific stop words list curated based on the frequency of occurrence of these words within a social network. We apply a number of rule based pattern matching methods to eliminate emoticons, slangs and similar variants that are typically popular on Facebook and Twitter. Profanity and adult content being prevalent on social networks, we use an adult and spam classifier to filter inappropriate social fragments. Additionally, we use an editorially defined profanity lookup to eliminate certain social network specific spam and adult content.

4.2 Features

We extract N-grams from the processed text from each document d_j, which constitute the candidates for social signatures for the associated URL u_j. A score is calculated for each candidate social signature using a model that computes a weighted combination of the following features:

1. Term frequency (tf) for a N-gram with respect to a particular document, given by the number of times the N-gram appears in that document. This feature captures the number of times users used this N-gram in their post and it follows that if many users are heavily using a particular N-gram in their social fragment, it is related with the relevance to the associated URL.
2. Document frequency (df) for a N-gram, given by the number of documents in which N-gram occurs. This feature measures if the specific N-gram is popular globally or just in the context of the particular URL, the N-grams with very high document frequency are less likely to be conveying information unique to a particular URL.
3. Term frequency-Inverse document frequency (tf-idf) of the N-gram, calculated as

$$\text{tf-idf} = 1 + \log(tf) * \log\left(\frac{n}{df}\right)$$

where n is the number of documents. This feature captures the normalized relevance of the N-gram for a particular URL.
4. Product of the tf-idf of each unigram comprising the N-gram. This feature tends to measure the strength of each of the unigrams comprising the N-grams. Higher order N-grams typically comprise of unigrams with high df; in spite of a high tf-idf measure for the N-gram, such a social signature is not meaningful and complete in describing the associated URL. Thus, this feature rewards the N-grams containing high tf-idf unigrams and penalizes N-grams comprising low tf-idf unigrams.
5. Local affinity of the N-gram, calculated as the ratio of the tf of the N-gram and the maximum tf among the unigrams comprising the N-gram. This feature measures the likelihood of co-occurrence of unigrams and N-gram. For example, this feature, assigns a high weight to N-grams such as “San Francisco” and “Asiana Airlines”, where the occurrence of any unigram within the N-gram is highly correlated with the occurrence of the complete N-gram and very low weight to N-grams like “Francisco crash” and “Airlines flight”.
6. Global affinity of the N-gram, calculated as the ratio of the df of the N-gram and the maximum df among the unigrams in the N-gram. This feature measures the co-occurrence affinity of the unigrams, similar to local affinity as described above but in the global context based on the language information gained based on all the documents. Thus, this feature reinforces the strength of the affinity of unigrams not limiting to the knowledge derived from the social anchor text for a specific URL.

We score each candidate for a social signature using a weighted linear model that combines the features mentioned above. The N-gram score pairs having a score greater than the threshold of 0.1% of the maximum score for a specific URL comprise the social signature for that URL. Our approach ensures that we make only a constant number of passes over the data. Since all of the features are count based and can be parallelized very well in a distributed framework, this method is very efficient.

4.3 Implementation Details

All the algorithms were implemented in the SCOPE language [5] and run over a large distributed computing cluster. We deploy a production pipeline which fetches the raw social feed and performs all the processing and computation of social signatures in the distributed computing cluster. Every Facebook post and the associated comment is mined to select the ones associated with a URL. We perform similar steps for Twitter.

All the classifiers and data processing operations are performed on the raw tweets and Facebook posts/comments to select relevant content. The social signatures computation is performed over the processed dataset. As explained in the previous subsection, all the features used in the model for computing social signatures are count based and hence can be parallelized extremely well over a large number of machines in a distributed computing cluster. The resulting pipeline is efficient and is used to generate social signatures in a production environment.

5. EXPERIMENTS AND EVALUATION

In this section we present the results of the experiments we performed using the datasets already introduced.

5.1 Social Signatures Computation

As mentioned earlier, the links dataset consists of 931M Twitter URLs and 116M Facebook unique candidate URLs for which we attempted to compute social signatures. In order to ensure that we got meaningful signatures, we imposed a minimum threshold on the number of times a bigram needs to be seen before we computed a score for it. Though this ensured that we had reasonable support for each candidate bigram, it also meant that we dropped URLs which did not have enough text around them. After processing data, we ended up with 7M unique URLs for the Twitter URLs having at least one social signature. Table 1 shows the percentage of Twitter URLs for which we were able to compute social signatures broken down by the number of times the URL was originally shared.

For the Facebook dataset, 4.5M URLs ended up with atleast one social signature. Table 2 shows the percentage of Facebook URLs for which we were able to compute social signatures broken down by the number of times the URL was originally shared.

URL frequency	URLs with social signatures	URLs in the original dataset	% of URLs with social signatures
1 – 10	77K	915M	0.00841%
10 – 100	6M	14.7M	41.02%
100 – 1K	800K	1.2M	64.48%
1K– 10K	51.5K	76.5K	67.36%
10K – 100K	2.6K	3K	86.83%
100K and above	164	180	91.11%

Table 1: Percentage of Twitter URLs with social signatures broken down by the number of times these URLs were shared.

URL Frequency	URLs with social signatures	URLs in the original dataset	% of URLs with social signatures
1 – 10	2.7M	112M	2.40%
10 – 100	1.5M	3.7M	40.11%
100 – 1K	270K	352.6K	76.64%
1K– 10K	23K	25.6K	89.06%
10K and above	941	1.2K	79.21%

Table 2: Percentage of Facebook URLs with social signatures broken down by the number of times these URLs were shared.

As illustrated in Table 1 and Table 2, we see that as the number of times a URL is shared increases, the possibility of it having a social signature also increases. Figure 8 shows the average number of signatures for a URL as a function of the number of times that URL is shared.

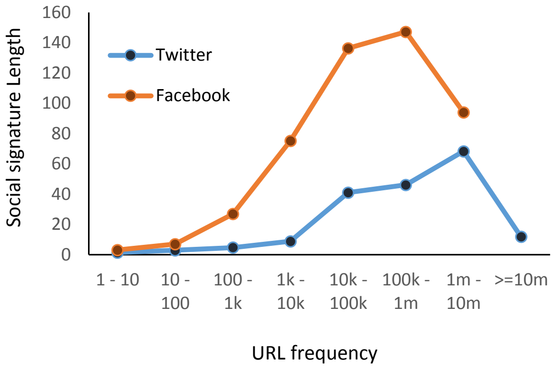


Figure 8: Average number of social signatures as a function of the number of times the URL is shared.

Interestingly the number of social signatures for a URL increases as the number of times a URL is shared. The number however falls for the URLs shared more than 10 million times. A closer analysis showed that there are very few URLs that fall in this bucket and these URLs were mainly related to domains such as `unfollowers.me` and `justunfollow.com` and thus contained terms that were considered to be stop words for the Twitter domain (e.g., Unfollow, followers retweet, etc.). As a result a lot of these terms did not have a score above the required threshold. Similar pattern was observed for Facebook where the URLs shared more than 1 million times corresponded to specific apps that are heavily popular on Facebook.

5.2 Comparison with the Web

In this section we compare the social signatures for a URL with the textual content of the web page. The difference between the web page content and the social signatures represents the new information for the URL gained from the social anchor text. We determine the difference between the N-grams in the social signature in comparison to the N-grams found in the web page text. Specifically, a given social signature is considered to convey new information about the URL if neither of the unigrams in the social signature are present in the web page content. If the web page contains both the tokens in a N-gram social signature as distant words in the web page text, then the social signature is not considered to be contributing new information.

Web page content for the Facebook and Twitter URLs is extracted as described in Section 3. This reduces the original shared links dataset to a subset of 368K URLs for Facebook and 1M URLs for Twitter. Figure 9 illustrates the comparison of social signatures with the web content for this subset of URLs. The dotted lines in the plot correspond to the average number of social signatures as a function of URL frequency based on the complete set of social signatures. Social signatures with N-grams not present in the web content, indicated by the solid lines is measured as the average number of social signatures providing new information about the URL compared to the web page content. The results suggest that the social signatures provide new information about the URL and the difference with the web content increases as URL shared frequency increases. It is interesting to note that the social anchor text for both Facebook and Twitter follow similar trends. The number of social signatures and the difference with the web content is higher for Facebook URLs in comparison to the Twitter URLs. The length of a tweet is limited by 140 characters whereas the social anchor text around a URL shared on Facebook is significantly more verbose, this explains the difference observed.

We analyze the coverage of URLs where social signatures provide new information compared to the web content. Figure 10 displays the percentage of Facebook and Twitter

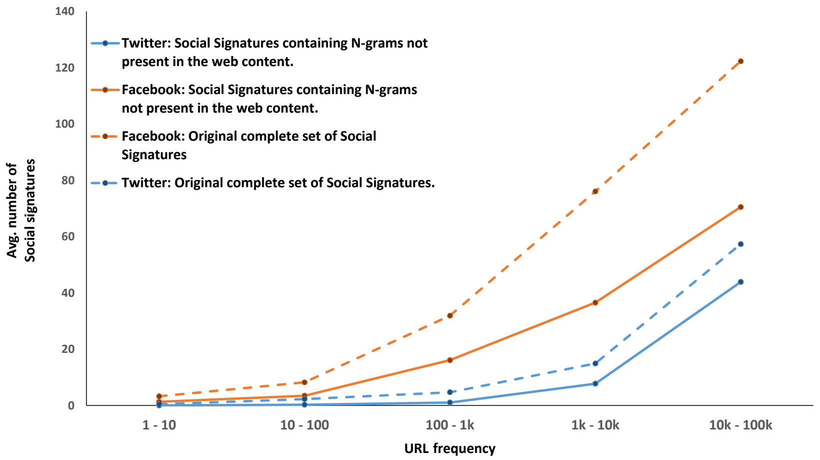


Figure 9: Comparison of the social signatures for a URL with the web page content as a function of the number of times the URL is shared.

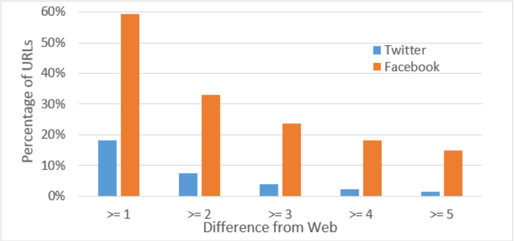


Figure 10: Distribution of URLs providing new insights compared to the web page content as a function of varying number of social signatures.

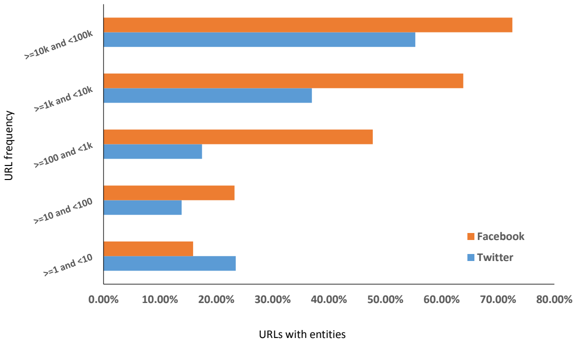
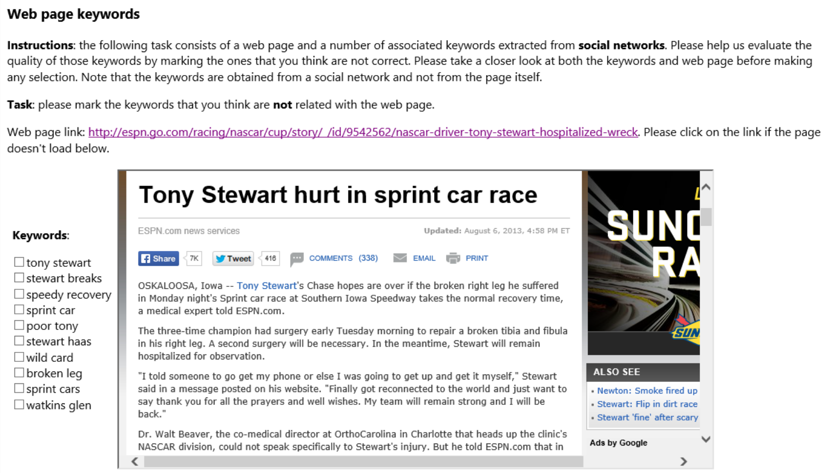


Figure 11: Distribution of URLs where new social insights include entities as a function of the number of times the URL is shared.

URLs providing new information as a function of varying number of social signatures. It is observed that 59.36% of Facebook URLs and 18.15% of Twitter URLs gain new information from social signatures. For 14.58% of Facebook URLs and 1.54% of Twitter URLs number of social signatures providing new information compared to the web content is at least 5 social signatures. As outlined in Section 3 it is important to note that majority of the URLs are shared 1-10 times and as illustrated in Figure 5 the number of social signatures for this set of URLs is extremely low due to lack of sufficient social anchor text. However, in Figure 7 the distribution does not differentiate amongst different shared frequency of URLs. The number of URLs gaining new information from social anchor text appear low because of the bias by URLs with low share frequency. Analyzing the social signatures for Facebook URLs, we observe that a significantly larger fraction of Facebook URLs gain new information from social signatures, compared to Twitter.

To assess the content of the N-grams providing new information compared to the web page, we identify the named entities in the social signatures. We use an internal named entity tagger trained on social data to recognize the entities in the social signatures that provide new information. Tagged entities correspond to four types, namely: `PERSON`, `LOCATION`, `ORGANIZATION`, and `PRODUCTS`. Figure 11 provides a distribution of the URLs containing named entities in the new information gained from social signatures, by considering only the URLs where social signatures provide new information. As displayed in the figure, we see that a large fraction of URLs gain new entities from social signatures which are missing from the original web content. Also, this fraction further increases with the increase in URL share frequency. Figure 12 further illustrates the distribution of these URLs across different entity types. The entity type `PERSON` is the dominant entity on both Facebook and Twitter, followed by `LOCATION` on Facebook and `ORGANIZATION` on Twitter.



Web page keywords

Instructions: the following task consists of a web page and a number of associated keywords extracted from **social networks**. Please help us evaluate the quality of those keywords by marking the ones that you think are not correct. Please take a closer look at both the keywords and web page before making any selection. Note that the keywords are obtained from a social network and not from the page itself.

Task: please mark the keywords that you think are **not** related with the web page.

Web page link: http://espn.go.com/racing/nascar/cup/story/_/id/9542562/nascar-driver-tony-stewart-hospitalized-wreck. Please click on the link if the page doesn't load below.

Keywords:

□ tony stewart
□ stewart breaks
□ speedy recovery
□ sprint car
□ poor tony
□ stewart haas
□ wild card
□ broken leg
□ sprint cars
□ watkins glen

Tony Stewart hurt in sprint car race

ESPN.com news services

Updated: August 6, 2013, 4:58 PM ET

OSKALOOSA, Iowa -- Tony Stewart's Chase hopes are over if the broken right leg he suffered in Monday night's Sprint car race at Southern Iowa Speedway takes the normal recovery time, a medical expert told ESPN.com.

The three-time champion had surgery early Tuesday morning to repair a broken tibia and fibula in his right leg. A second surgery will be necessary. In the meantime, Stewart will remain hospitalized for observation.

"I told someone to go get my phone or else I was going to get up and get it myself," Stewart said in a message posted on his website. "Finally got reconnected to the world and just want to say thank you for all the prayers and well wishes. My team will remain strong and I will be back."

Dr. Walt Beaver, the co-medical director at OrthoCarolina in Charlotte that heads up the clinic's NASCAR division, could not speak specifically to Stewart's injury. But he told ESPN.com that in

ALSO SEE

- Newton: Smoke fired up
- Stewart: Flip in dirt race
- Stewart 'fine' after scary

Figure 13: HIT task design. Workers need to mark which N-grams (keywords in the task) are not related to the web page that is presented.

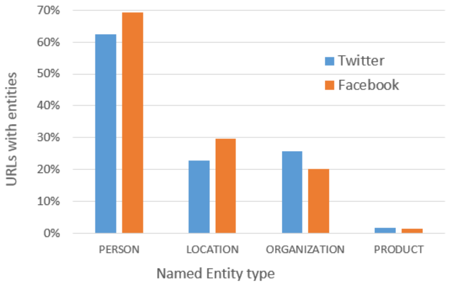


Figure 12: Distribution of entity types.

5.3 Human Evaluation

The last phase of our evaluation methodology includes a quality assessment of the social signatures using human computation. To collect assessments, we follow a similar approach to the ESP Game [19] where players have to agree on image labels. Similarly, we expect workers to agree on the good and bad labels. We designed and implemented a task that requires a worker to mark which N-gram is not representative or descriptive enough. We also provided a link to the web page in case the content was not visible in the human intelligence task. Figure 13 shows a screenshot of such task.

We sample a set of 400 random URLs from the bucket 1K-10K in terms of URL frequency. The task was created in Mechanical Turk, using 3 workers per URL, with a payment of $0.04 per HIT. The average number of N-grams per URL on this data set is 9.62 with a standard deviation of 1.67.

As expected, this is a very subjective task that requires understanding the document and the potential social context. We assume that the list of N-grams in the signature presented to the user is of good quality and we only remove an item from such list if there is consensus from the workers that the N-gram is bad.

For our evaluation purposes we cast the problem as a ranked list binary relevance assessment task. As input we have the social signatures, the N-gram list, ranked by relevance. Each worker assesses if a particular keyword (N-gram) is not relevant to the document by marking the appropriate check-box. That is, a check means not relevant (0) and a no-check means relevant (1). The final label is computed using majority vote. We illustrate the process using the following URL-social signature pair: `(http://act.freepress.net/sign/journ_press_intimidation, {united states, journalisms role, national security, civil liberties, totalitarian regimes, edward snowden, leaders accountable, specific case, individual journalists, police state})`. Table 4 shows the assessments by the three workers and final assigned label using majority vote.

Rank	Keyword	W1	W2	W3	Label
1	united states	1	1	1	1
2	journalisms role	1	1	1	1
3	national security	1	1	1	1
4	civil liberties	1	1	1	1
5	totalitarian regimes	1	0	1	1
6	edward snowden	1	1	1	1
7	leaders accountable	1	0	0	0
8	specific case	0	1	0	0
9	individual journalists	0	0	0	0
10	police state	1	0	0	0

Table 4: Relevance assessments provided by each worker and final aggregated label using majority vote.

In Table 3 we present a number of examples of URLs, associated social signatures, and defect rate. It is very interesting to see how the signatures capture what the activity around that page seems to be, without judging the sentiment of a particular social network.

We first look at label distribution from workers as follows. Each worker can mark on the task a number of N-grams as bad. If we average the number of incorrect N-grams per URL, it gives us an idea of the defect rate of the social

Source type	**URL**	**Social signature**	**Defect rate**
Text and video	`www.undergroundhealth.com/coke-is-blatantly-lying-about-aspartames-dangers`	coca cola, peer reviewed, corn syrup, side effects, diet coke, artificial sweeteners, fructose corn, bottom line, high fructose, diet soda	0.15
Text	`www.cnn.com/2013/08/28/world/europe/new-chemical-element/index.html`	element 115, miley cyrus, element discovered, stable element, tony stark, bang theory, area 51, shower curtain, heavy elements, science geek	0.3
Video	`www.godvine.com/Fatherless-Bride-Does-the-Most-Touching-Thing-at-Her-Wedding-3808.html`	brought tears, young woman, aisle, wedding, walk, dad, touching, bride, father, voice	0.2
Video	`vimeo.com/70994185`	goose bumps, air force, aerial footage, plane porn, aerial shots, full screen, absolutely stunning, high quality, top gun, red epic	0.1
Text	`downtrend.com/brian-carey/heres-3-constitutional-rights-liberals-admit-they-want-to-take-away`	united states, bear arms, 2nd amendment, regulated militia, constitutional rights, civil liberties, religious freedom, fifth amendment, due process, second amendment	0.2
Text	`www.mrconservative.com/2013/05/17955-islamberg-usa-a-mulsim-only-town-in-new-york`	united states, ruby ridge, mr conservative, muslim brotherhood, tea party, politically correct, jihad training, upstate ny, radical islam, homeland security	0.05
Text and video	`www.upworthy.com/forget-everything-you-learned-in-economics-you-were-totally-lied-to`	bottom line, conventional economics, keynesian economics, david suzuki, ecological economics, environmental economics, pay attention, economics courses, economics class, economics teacher	0.13
Text	`www.reagancoalition.com/articles/2013/20130726006-chaplain-censor.html`	founding fathers, politically correct, political correctness, military chaplain, jesus christ, military chaplains, dwight eisenhower, religious beliefs, air force, lt col	0.3
Text	`www.newyorker.com/online/blogs/closeread/2013/07/what-should-trayvon-martin-have-done.html`	trayvon martin, jelani cobb, zimmerman verdict, zimmerman trial, george zimmerman, zimmerman instigated, united states, zimmerman initiated, punched zimmerman, confronted zimmerman	0.35

Table 3: Examples of URL-social signature pairs along with they their defect rate. Some of the topics are controversial and it is expected to see a lot of discussion. URLs that have a video tend to draw a lot of attention.

signature. We define the defect rate, DR, per URL, as

$$DR_{url} = \frac{\#\text{incorrect n-grams}}{\#\text{n-grams}}$$

Figure 14 shows the results using raw answers from workers. While this result does not include any voting yet, it gives an early impression regarding quality. We observe a low defect for most of the URLs. We can classify the quality problems into two categories: 1) conflation (e.g., `(romo fans, romo fan)`, `(sarah palin, sara palin)`) and 2) incorrect English language detection making difficult for workers to assess the content. We also noticed that some of the content has a temporal context so it desirable to assess as soon as social signatures can be computed.

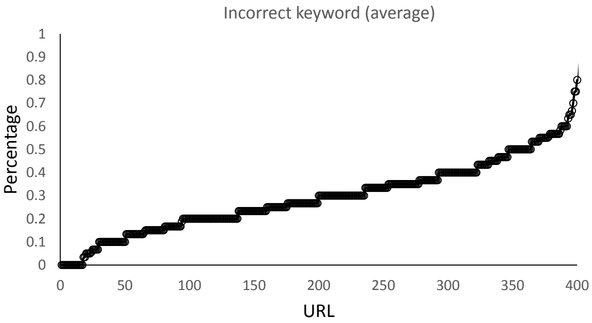


Figure 14: Average defect rate per URL using raw answers from workers.

We pick Average Precision (AP) as a measure because it is top-heavy, that is it is sensitive to changes near the top ranks, and because the order does matter. It is better to submit more certain N-gram first, followed by N-grams we are less sure about. We define AP as follows:

$$AP = \frac{1}{R} \sum_{r} I(r) \frac{C(r)}{r}$$

where r is rank, $I(r)$ is 0 if the N-gram at rank r is not relevant and 1 if the N-gram is relevant, and $C(r)$ is the number of relevant n-grams within r. R is the number of relevant N-grams.

We use mean average precision (MAP) to compute the precision of the signatures among the three workers, that is

$$MAP_{url} = \frac{AP}{3}$$

The results are presented in Figure 15, this time including the voting outcome.

6. APPLICATIONS

So far, we presented a technique called *social signatures* to extract *temporally and socially salient* keywords from social media content associated with a link. These signatures can augment a given link with new metadata and be used in a number of different scenarios.

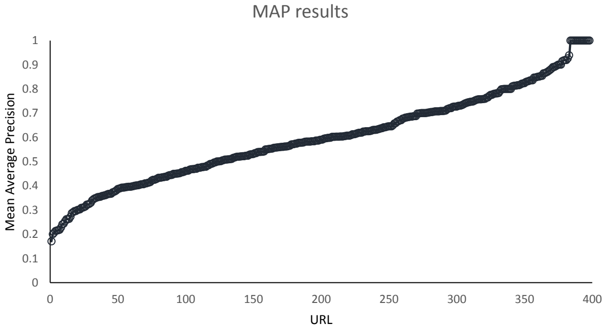


Figure 15: MAP results using aggregated votes.

Details of the utility of social signatures on those scenarios and other specific metrics are left out due to proprietary information. That said, we outline the following applications.

1. New ranking signals for search engine relevance.
2. Metadata for enhancing image and video search.
3. Faster indexing of pages with high activity by providing signals to a crawler regarding queue prioritization.
4. Taxonomy generation and discovery of new vocabulary not restricted to content directory structures like ODP (Open Directory Project).
5. New browsing and discovery experiences based on alternatives mechanisms for organizing shared content.
6. Recommendations based on similarity between user profiles and signatures.
7. Exploratory and faceted search applications for social archives similar to the work reported in [1].

7. CONCLUSIONS AND FUTURE WORK

We described the design and implementation of our algorithms at scale along with the data set characteristics that we used to showcase the evaluation results. While we showed results using English-only content, the techniques presented are multilingual. Our techniques and data gathering were presented exhaustively so it should be possible to replicate the results with smaller samples.

We also included a detailed analysis of social signatures related to URLs shared on Twitter and Facebook. Our techniques should work with other types of social networks that rely on sharing content with text annotations like Google+, Instagram, Foursquare or LinkedIn.

As part of our data analysis, we provided some insights into how this content differs from what is found on the Web and conducted a crowdsourcing-based evaluation that supports our thesis that annotations obtained from social media content add contextually relevant information improving the quality of the recommendations.

Future work includes the analysis of the social anchor text from additional networks as well as the comparison and differences in the nature of the social signatures obtained from different networks. We see the detailed analysis of the use of social signatures to predict high frequency queries as a promising direction for this type of work.

8. REFERENCES

[1] Omar Alonso and Kartikay Khandelwal. Kondenzer: Exploration and visualization of archived social media. In *Proceedings of ICDE*, 2014.

[2] Einat Amitay, Adam Darlow, David Konopnicki, and Uri Weiss. Queries as anchors: selection by association. In *Proceedings of Hypertext*, pages 193–201, 2005.

[3] Peter Anick. Exploiting anchor text as a lexical resource. In *LREC*, 2004.

[4] Oisı Boydell and Barry Smyth. Social summarization in collaborative web search. *Information processing & management*, 46(6):782–798, 2010.

[5] Ronnie Chaiken, Bob Jenkins, Per-Åke Larson, Bill Ramsey, Darren Shakib, Simon Weaver, and Jingren Zhou. Scope: Easy and efficient parallel processing of massive data sets. *PVLDB*, 1(2):1265–1276, August 2008.

[6] Nadav Eiron and Kevin McCurley. Analysis of anchor text for web search. In *Proceedings of SIGIR*, pages 459–460, 2003.

[7] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of CIKM*, pages 1625–1628, 2010.

[8] Atsushi Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. In *Proceedings of WWW*, pages 337–346, 2008.

[9] Michael Gamon, Tao Yano, Xinying Song, Johnson Apacible, and Patrick Pantel. Understanding document aboutness step one: Identifying salient entities. *MSR-TR-2013-73*, 2013.

[10] Carolin Gerlitz and Anne Helmond. The like economy: Social buttons and the data-intensive web. *New Media Society*, 15:1348–1365, 2013.

[11] Chia-Jung Lee and Bruce Croft. Incorporating social anchors for ad hoc retrieval. In *Proceedings of OAIR*, pages 181–188, 2013.

[12] Donald Metzler, Jasmine Novak, Hang Cui, and Srihari Reddy. Building enriched document representations using aggregated anchor text. In *Proceedings of SIGIR*, pages 219–226, 2009.

[13] Gilad Mishne and Jimmy Lin. Twanchor text: a preliminary study of the value of tweets as anchor text. In *Proceedings of SIGIR*, pages 1159–1160, 2012.

[14] Aditi Muralidharan, Zoltan Gyongyi, and Ed Chi. Social annotations in web search. In *Proceedings of SIGCHI*, pages 1085–1094, 2012.

[15] Michael Noll and Christoph Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Proceedings of Web Intelligence*, volume 1, pages 640–647, 2008.

[16] Patrick Pantel, Michael Gamon, Omar Alonso, and Kevin Haas. Social annotations: Utility and prediction modeling. In *Proceedings of SIGIR*, pages 285–294, 2012.

[17] Seung-Taek Park, David Pennock, C Lee Giles, and Robert Krovetz. Analysis of lexical signatures for finding lost or related documents. In *Proceedings of SIGIR*, pages 11–18, 2002.

[18] Stéphane Raux, Nils Grünwald, and Christophe Prieur. Describing the web in less than 140 characters. In *Proceedings of ICWSM*, 2011.

[19] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of SIGCHI*, pages 319–326, 2004.

[20] Mingfang Wu, David Hawking, Andrew Turpin, and Falk Scholer. Using anchor text for homepage and topic distillation search tasks. *Journal of the American Society for Information Science and Technology*, 63(6):1235–1255, 2012.

[21] Bo Zhou, Yiqun Liu, Min Zhang, Yijiang Jin, and Shaoping Ma. Incorporating web browsing activities into anchor texts for web search. *Information Retrieval*, 14(3):290–314, 2011.