Correlation of Term Count and Document Frequency for Google N-Grams

Martin Klein and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk VA 23529

{mklein, mln}@cs.odu.edu

Abstract. For bounded datasets such as the TREC Web Track (WT10g) the computation of term frequency (TF) and inverse document frequency (IDF) is not difficult. However, when the corpus is the entire web, direct IDF calculation is impossible and values must instead be estimated. Most available datasets provide values for *term count (TC)* meaning the number of times a certain term occurs in the entire corpus. Intuitively this value is different from *document frequency (DF)*, the number of documents (e.g., web pages) a certain term occurs in. We investigate the relationship between TC and DF values of terms occurring in the Web as Corpus (WaC) and also the similarity between TC values obtained from the WaC and the Google N-gram dataset. A strong correlation between the two would gives us confidence in using the Google N-grams to estimate accurate IDF values which for example is the foundation to generate well performing lexical signatures based on the TF-IDF scheme. Our results show a very strong correlation between TC and DF within the WaC with Spearman's $\rho \geq 0.8$ ($p \leq 2.2 \times 10^{-16}$) and a high similarity between TC values from the WaC and the Google N-grams.

1 Introduction and Motivation

In information retrieval (IR) research term frequency (TF) - inverse document frequency (IDF) is a well known and established term weighting concept. TF-IDF extracts the most significant terms from textual content while also dismissing more common terms such as stop words. It is often used for term weighting in the vector space model as described by Salten et al. [15]. It further can be used to generate lexical signatures (LSs) of web pages as shown in [14, 13, 6, 10, 18].

The computation of TF values for a web page is straight forward since we can simply count the occurrences for each term within the page. The computation of IDF values however is more complex. Two values are mandatory:

1. the overall number of documents in the corpus and
2. the number of documents a term appears in.

Since both values are unknown when the entire web is the corpus, accurate IDF computation for web pages is impossible and values need to be estimated. For

simplicity we call the second mandatory value *document frequency (DF)* which is different from *term count (TC)*. The following simple example to illustrate the difference between TC and DF. Let us consider a corpus of 4 documents $D = d_1...d_4$ where each document contains the title of a song by The Beatles. Table 1 shows the documents as well as the TC and DF values of all terms occurring in our small sample corpus. We can see that the values are identical for the majority of the terms (8 out of 10). The general concept behind our earlier

Table 1. *TC-DF* Comparison Example

$d_1 = Please\ Please\ Me$	$d_3 = All\ You\ Need\ Is\ Love$
$d_2 = Can't\ Buy\ Me\ Love$	$d_4 = Long,\ Long,\ Long$

Term	All	Buy	Can't	Is	Love	Me	Need	Please	You	Long
TC	1	1	1	1	2	2	1	2	1	3
DF	1	1	1	1	2	2	1	1	1	1

research introduced in [10] is using LSs to (re-)discover missing web pages. That is, once a web user experiences a 404 "Page not found" error we query search engine caches and the Internet Archive (IA) for copies of the missing page. In case these (old) copies are not sufficient for the user we generate LSs from the obtained copies of the missing web page and use them to query search engines for the missing page at its new location and alternative pages that potentially also satisfy the user's information need. Since our LSs are generated from live web pages (even though taken from search engine caches or the IA) and they have to be generated in real time the question arises how to estimate DF values.

Corpora containing the textual content of web pages can be used to compute or estimate DF values. These corpora are generally considered a representative sample for the Internet [16] but have also been found to be somewhat dated [1]. The *TREC Web Track* is probably the most common corpus and has, for example, been used in [18] for IDF estimation. The *British National Corpus (BNC)* [11], as another example, has been used in [17]. These corpora provide precise IDF values since the total number of documents is known and DF values can also be determined by simply parsing the single resources and therefore both mandatory values for the IDF computation are given. Both corpora are not freely available and hence we turn our focus to the *Google N-grams* [4] published in 2006. This corpus is based on the Google index and therefore provides a powerful alternative to the corpora mentioned above. The N-grams unfortunately only provide TC values of all terms (n-tokens) and Google does not intend to publish the DF values any time soon. Therefore we are motivated to investigate the correlation between TC and DF values. In case of a positive outcome we would be able to use the TC values provided by the N-grams to estimate accurate IDF values for our LSs. The *Web as Corpus kool ynitiative (WaCky)*[1] provides the

[1] http://wacky.sslmit.unibo.it/doku.php

WaC with no charge for researchers. The corpus provides TC values of all terms it contains and their DF values can also be determined.

The contribution of this paper is the investigation of the relationships

1. between TC and DF values within the WaC and
2. between WaC based TC and Google N-gram based TC values.

2 Related Work

Zhu and Rosenfeld [19] used Internet search engines to obtain estimates for DF values of unigrams, bigrams and trigrams. They plotted the obtained phrase count (comparable to what we call TC) and web page count (our DF) and were able to apply a log-linear regression function to all three n-gram cases which implies a strong correlation between the obtained TC and DF values. Zhu and Rosenfeld also found that the choice of one particular search engine did not matter much for their results.

Keller and Lapata [8] also used Internet search engines to obtain DF values for bigrams. They compare these values to corpus frequencies (comparable to our TC) obtained from the BNC and the North American News Text Corpus (NANTC). Despite significant differences between the two corpora, Keller and Lapata found a strong correlation between the web based values (DF) and the values obtained from the two text corpora (TC). The main application Keller and Lapata see for their results is estimating frequencies for bigrams that are missing in a given corpus.

Nakov and Hearst [12] approach to give a justification for using the search engine result count (DF) as estimates for n-gram frequencies (which can be TC). They chose the noun compound bracketing problem (which has traditionally been addressed by using n-gram frequencies) to demonstrate their results. They found that the n-gram count from several Internet search engines differs and these differences are measurable but not statistically significant. They come to the conclusion that the variability over time and across different search engines represented by the obtained n-gram frequencies does not impact the results of a specific natural language processing task.

All these studies have two things in common: 1) they all show a strong correlation between DF and TC values and 2) they use DF estimates from search engines and compare it to TC values from either conventional corpora or the web as well. This is where our approach is different since we use TC values from well established text corpora and show the correlation to measured DF values obtained from these corpora.

Sugiyama et al. [18] used the TREC-9 Web Track dataset [7] to estimate IDF values for web pages. The novel part of their work was to also include the content of hyperlinked neighboring pages in the TF-IDF calculation of a centroid page. They show that augmenting TF-IDF values with content of in-linked pages increases the retrieval accuracy more than augmenting TF-IDF values with content from out-linked pages. Their research is based on the idea

that the content of a centroid web page is often related to the content of its neighboring pages which has also been shown in [2] and [3].

Phelps and Wilensky [14] proposed using the TF-IDF model to generate LSs of web pages and introduced "robust hyperlinks", an URL with a LS appended. Phelps and Wilensky conjectured if an URL would return a HTTP 404 error, the web browser could submit the appended LS to a search engine to either find a copy of the page at a different URL or a page with similar content compared to the missing page. Phelps and Wilensky did not publish details about how they determined IDF values but stated that the mandatory figures can be taken from Internet search engines. That implies the assumption that the index of a search engine is representative for all Internet resources. However, they do not publish the value they used for the estimated total number of documents on the Internet.

3 Experiment Design

The WaC provides a frequency list of all unique terms in the corpus (lemmatized and non-lemmatized) and their TC value. The document boundaries in the corpus are given, hence we can compute the DF values for all terms. Since we are interested in generating TF-IDF values for web pages and feeding them back into search engines we only use non-lemmatized terms. Conducting a similar set of experiments using the lemmatized terms remains for future work. We rank both lists in decreasing order of their TC and DF values and investigate the relationship between the rankings by computing:

1. Spearman's ρ and Kendall τ of the ranked list of terms (results are shown in Section 4.1) and
2. the frequency of TC/DF ratio of all terms (results in Section 4.2).

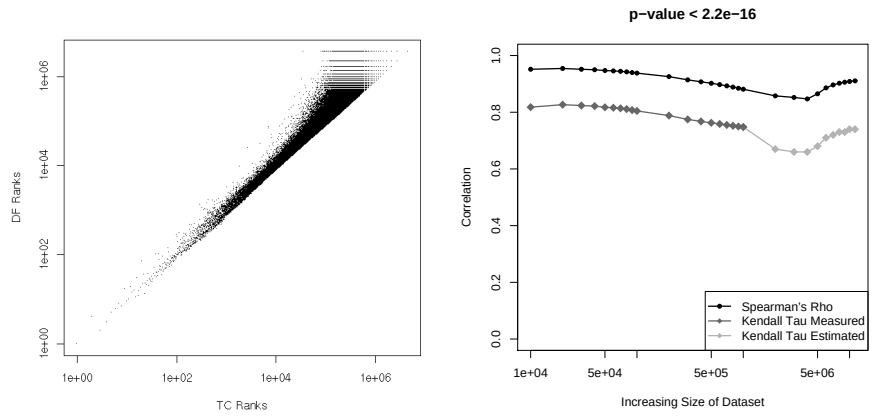
The results of the comparison between TC frequencies of the WaC and the N-gram corpus are shown in Section 4.3.

4 Experiment Results

4.1 Correlation Within the WaC

Figure 1(a) shows (in loglog scale) the TC and DF ranks of all terms from our WaC dataset. The x-axis represents the TC ranks and the y-axis the corresponding DF ranks. We see the majority of the points within a diagonal corridor which indicates a high similarity between the rankings since two identical lists would be displayed as a perfect diagonal line.

Figure 1(b) shows the measured and estimated correlation between TC and DF values in the WaC dataset. The increasing size of the dataset, meaning the increasing list of terms, is shown on the x-axis. The solid black line displays the Spearman's ρ values. The value for ρ at any size of the dataset is above 0.8 which indicates a very strong correlation between the rankings. The results are



(a) Ranks in Log-Log Scale (b) Correlation in Semi-Log Scale

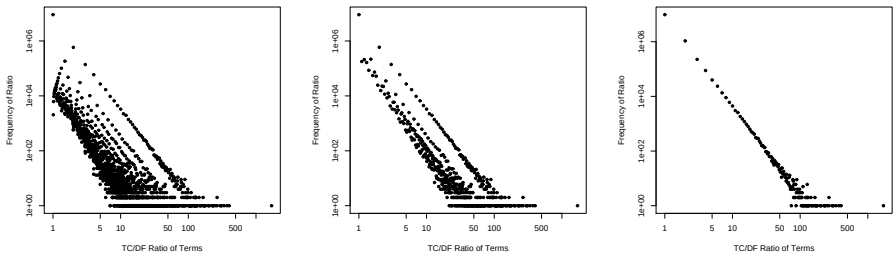
Fig. 1. Ranks and Measured and Estimated Correlation between Term Count and Document Frequency in the WaC dataset

statistically significant with a p-value of 2.2×10^{-16}. The green solid line shows the computed Kendall τ values for the top $1,000,000$ ranks and the dotted blue line represents the estimated τ values for the remaining ranks. We again find a strong correlation with computed τ values between 0.82 and 0.74 and estimated τ values of at least 0.66. We did not compute τ for greater ranks since it is a very time consuming operation. Gilpin [5] provides a table for converting τ into ρ values. We use this data to estimate our τ values. Even though the data in [5] is based on τ values computed from a dataset with bivariate normal population (which we do not believe to have in the WaC dataset), it supports our measured values.

4.2 Term Count - Document Frequency Ratio in the WaC

Another way of displaying the correlation between the two rankings is plotting the TC/DF ratios of all terms. For two ranked lists which are perfectly correlated the ratio for all list items would be equal to 1. Figure 2 shows (in loglog scale) the frequency of all ratios and confirms the dominance of the "perfect ratios". Figure 2(a) shows the distribution of TC/DF ratios with values rounded after the second decimal and Figure 2(b) shows the ratios rounded after the first decimal. It becomes obvious that the vast majority of the ratio values are close to 1. The visual impression is supported by the computed mean value of 1.23 with a standard deviation of $\sigma = 1.21$ for both, Figure 2(a) and 2(b). The median of ratios is 1.00 and 1.0 respectively. Figure 2(c) shows the distribution of TC/DF ratios rounded as integer values. It is consistent with the pattern of Figures 2(a) and 2(b) and the mean value is equally low at 1.23 ($\sigma = 1.22$). The median here is also 1. Figure 2 together with the computed mean and median values accounts

for another solid indicator for the strong correlation between TC and DF values within the corpus.



(a) Two Decimals (b) One Decimal (c) Integer Values

Fig. 2. Frequency of TC/DF Ratios in the WaC - Rounded

4.3 Similarity Between WaC and N-gram TC Values

As mentioned in Section 1, after showing the correlation between TC and DF values within the WaC, it is our goal to investigate the similarity between the TC values available from both corpora, WaC and the Google N-grams. Since both corpora are based on different sources and the N-gram dataset was generated from a much greater set of documents a direct comparison of intersecting terms could be misleading. However, a comparison of the frequency of all TC values of the two corpora will give an indication of the similarity of the two datasets. Figure 3 displays (in loglog scale) these frequencies of unique TC values from both corpora. Visual observation of the figure confirms the intuition that the distribution of TC values in both corpora is very similar. It is our assumption that just the size of the Google N-gram corpus is responsible for the offset between the graphs. The graph further shows the TC threshold of 200 that Google applied while creating the N-gram dataset (unigrams occurring less than 200 times in their set of documents were dismissed).

Now, knowing that the TC values are very similar between the two corpora, we can say that the TC values provided by the Google N-gram dataset can be used to estimate accurate DF values. This result supports our research on generating accurate LSs by providing proper IDF values.

5 Conclusion and Future Work

We have shown a very strong correlation between the TC and DF ranks within the WaC with Spearman's $\rho \geq 0.8$ ($p \leq 2.2 \times 10^{-16}$). Our results further indicate a high similarity between TC values of the WaC and the Google N-gram

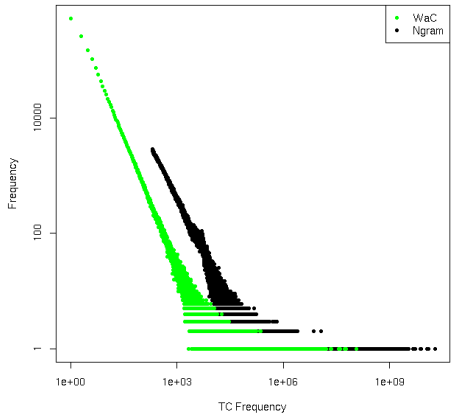


Fig. 3. Term Count Frequencies in the WaC and N-gram Corpus

corpus despite their difference in size. These results do not prove that all values correlated to TC (TF may be an example) can automatically be used as a replacement for DF but they give a strong indicator that the TC values gained from the Google N-grams (a recently generated, on web pages based corpus) are usable for the generation of accurate IDF values. In fact previous work [9] has shown a high similarity between LSs based on various sources such as the Google search engine and the Google N-grams. The computation of accurate IDF values now becomes more convenient and the resulting LSs of web pages still perform well in (re-)discovering the page when fed back into search engines.

In order to widen the spectrum of this work in the future we will use the TREC WT10g and the more recent TREC GOV collection and conduct experiments based on TC and DF values of these corpora. This work so far states that the correlation between TC and DF values is very high but further research will investigate whether we can determine a factor between the two values based on all available corpora. Further, this work does not guarantee a strong correlation between the raw TC and DF values. The effect of possibly inflated DF values for high, medium and low frequency terms remains for future work. Lastly, we used the non-lemmatized terms from the WaC and thus the correlation between the lemmatized ranked terms as well as the correlation of lists without stop words are still left to be investigated. We anticipate the correlation values to be slightly higher and will investigate the significance of this delta.

6 Acknowledgements

We thank the Linguistic Data Consortium, University of Pennsylvania and Google, Inc. for providing the "Web 1T 5-gram Version 1" dataset. We also thank the WaCky community for providing the ukWaC dataset. Further we would like to

thank Thorsten Brants from Google Inc. for promptly answering our emails and helping to clarify questions on the Google N-gram corpus.

References

1. W.-T. M. Chiang, M. Hagenbuchner, and A. C. Tsoi. The WT10G Dataset and the Evolution of the Web. In *Proceedings of WWW '05*, pages 938–939, 2005.
2. B. D. Davison. Topical locality in the web. In *Proceedings of SIGIR '00*, pages 272–279, 2000.
3. J. Dean and M. R. Henzinger. Finding Related Pages in the World Wide Web. *Computer Networks*, 31(11-16):1467–1479, 1999.
4. A. Franz and T. Brants. All Our N-Gram are Belong to You. `http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html`.
5. A. R. Gilpin. Table for Conversion of Kendall's Tau to Spearman's Rho Within the Context of Measures of Magnitude of Effect for Meta-Analysis. *Educational and Psychological Measurement*, 53(1):87–92, 1993.
6. T. L. Harrison and M. L. Nelson. Just-in-Time Recovery of Missing Web Pages. In *Proceedings of HYPERTEXT '06*, pages 145–156, 2006.
7. D. Hawking. Overview of the TREC-9 Web Track. In *NIST Special Publication 500-249: TREC-9*, pages 87–102, 2001.
8. F. Keller and M. Lapata. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29(3):459–484, 2003.
9. M. Klein and M. L. Nelson. A Comparison of Techniques for Estimating IDF Values to Generate Lexical Signatures for the Web. In *Proceedings of WIDM '08*, 2008.
10. M. Klein and M. L. Nelson. Revisiting Lexical Signatures to (Re-)Discover Web Pages. In *Proceedings of ECDL '08*, pages 371–382, 2008.
11. G. Leech, L. P. Grayson, and A. Wilson. Word Frequencies in Written and Spoken English: based on the British National Corpus. Longman, London, 2001.
12. P. Nakov and M. Hearst. A Study of Using Search Engine Page Hits as a Proxy for n-gram Frequencies. In *Proceedings of RANLP '05*, 2005.
13. S.-T. Park, D. M. Pennock, C. L. Giles, and R. Krovetz. Analysis of Lexical Signatures for Improving Information Persistence on the World Wide Web. *ACM Transactions on Information Systems*, 22(4):540–572, 2004.
14. T. A. Phelps and R. Wilensky. Robust Hyperlinks Cost Just Five Words Each. Technical Report UCB//CSD-00-1091, University of California at Berkeley, Berkeley, CA, USA, 2000.
15. G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.
16. I. Soboroff. Do TREC Web Collections Look Like the Web? *SIGIR Forum*, 36(2):23–31, 2002.
17. J. Staddon, P. Golle, and B. Zimny. Web based inference detection. In *USENIX Security Symposium*, 2007.
18. K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. In *Proceedings of HYPERTEXT '03*, pages 198–207, 2003.
19. X. Zhu and R. Rosenfeld. Improving Trigram Language Modeling with the World Wide Web. In *Proceedings of ICASSP '01*, pages 533–536, 2001.