Approximating Document Frequency with Term Count Values

Martin Klein, Michael L. Nelson
Old Dominion University, Department of Computer Science
Norfolk, VA 23529 USA

July 23, 2008

{mklein, mhn}@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 conduct a comparison study between TC and DF values within the Web as Corpus (WaC). We found a very strong correlation with Spearman’s $\rho \geq 0.8$ ($p \leq 0.005$) which makes us confident in claiming that for such recently created corpora the TC and DF values can be used interchangeably to compute IDF values. These results are useful for the generation of accurate lexical signatures based on the TF-IDF scheme.

1 Introduction and Motivation

In information retrieval (IR) research the term frequency (TF) - inverse document frequency (IDF) concept is well known and established to extract the most significant terms while dismissing the more common terms from textual content. It also has been used to generate lexical signatures (LSs) of web pages [1, 2, 3, 4, 5]. Such LSs can be used to (re-)discover missing web pages when fed back into search engine interfaces. The computation of TF values for a web page is straightforward since we can simply count the occurrences for each term within the page. Two values are mandatory for the IDF computation: the overall amount of documents in the corpus and the amount of documents a term appears in. We call the second value document frequency (DF). 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.

Various corpora containing web pages, their textual content and their in- and outlinks are available and can be used to estimate IDF values since they are considered a representative sample for the Internet [6]. The TREC Web Track is probably the most common corpus and has, for example, been used in [5] for IDF estimation. The British National Corpus (BNC) [7], as another example, has been used in [8]. Google published the N-grams [9] in 2006 and hence provides a powerful alternative source for TC values of terms extracted from web pages from the Google index. The Web as Corpus kool initiative (WaCky) provides the WaC corpus as an alternative with no charge for researchers. The problem with these corpora is that they do not provide DF values for the terms (or n-term tokens) they contain. We can count the total number of documents and therefore determine the DF values in case the corpus documents are provided along with the terms. Table 1 gives an overview of selected corpora and their characteristics. The first row indicates what kind of documents the corpus is based upon. The row Number of Documents shows the total number of documents the corpus consists of (or in the case of the Google N-grams the number of documents the corpus was generated from). This row also gives information about whether the documents of the corpus are
Table 1: Available Text Corpora Characteristics

| Corpus          | Google N-gram | TREC WT10g | BNC                      | WaC                      |
|-----------------|---------------|------------|--------------------------|--------------------------|
| Source          | Google indexed English language Web Pages | English language Web Pages | British English Texts (newspapers/journals, books), Transcripts of Verbal Language (meetings, radio shows) | uk-Domain Web Pages |
| Date            | 2006          | 1997       | 1994                     | 2006                     |
| Unique Terms    | > 13M         | 5.8M       | N/A                      | > 10M                    |
| Number of Documents | > 1B (Not Available) | 1.6M (Available) | 4,124                    | > 2.6M (Available)       |
| TC              | Available     | Not Available | Available from 3rd Party | Available                 |
| Freely Available | No*           | No         | No                       | Yes                      |

* A limited number of free copies of the corpus are available from the Linguistic Data Consortium, University of Pennsylvania

Table 2: TC-DF Comparison Example

| Term | All | Buy | Can’t | Is | Love | Me | Need | Please | You | My | Loving | Long |
|------|-----|-----|------|----|------|----|------|--------|-----|----|--------|------|
| TC   | 2   | 1   | 1    | 1  | 2    | 2  | 1    | 2      | 1   | 1  | 1      | 3    |
| DF   | 2   | 1   | 1    | 1  | 2    | 2  | 1    | 1      | 1   | 1  | 1      | 1    |

available. As mentioned above, recognizing the document boundaries within the corpus becomes necessary when computing IDF values.

The row TC indicates whether TC values of the corpus are available. The following simple example is to illustrate the difference between TC and DF. Let us consider a corpus of 5 documents \( D = d_1 \ldots d_5 \) where each document contains the title of a song by The Beatles:

\[
d_1 = \text{Please Please Me} \\
d_2 = \text{Can’t Buy Me Love} \\
d_3 = \text{All You Need Is Love} \\
d_4 = \text{All My Loving} \\
d_5 = \text{Long, Long, Long}
\]

Table 2 shows 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 example also shows that term processing such as stemming would have an impact on these numbers since Love and Loving are here treated as different terms.

Since we are interested in computing accurate IDF values for web page content it seems reasonable to chose a corpus that is based on textual content of web pages. Even though the TREC WT10g provides the documents and the corpus size seems sufficiently large, it has been shown to be somewhat dated [11].

We are interested in using the Google N-gram dataset as a corpus to generate accurate IDF values from
but unfortunately Google only provides $TC$ values. The WaC corpus in contrast provides both, $TC$ and $DF$ values and therefore we can:

1. establish a relationship between $TC$ and $DF$ values within the WaC
2. establish a relationship between WaC based $TC$ and Google N-gram based $TC$
3. and finally infer Google N-gram $DF$ from point 1 and point 2

This paper presents the preliminary results of the study and the results indicate that for sufficiently sized and recently generated corpora $DF$ values can be estimated from $TC$ values.

2 Related Work

2.1 Correlation between $DF$ and $TC$ Values

Zhu et al. [12] used an Internet search engine to obtain estimates for $DF$ values of n-grams. They used these values to estimate $TC$ values and compared those to $TC$ values from a 103 million word Broadcast News corpus which acted as their baseline. They found that the values are very similar and thus conclude that values obtained from the web are usable to estimate $TC$. Keller et al. [13] also used Internet search engines to obtain $DF$ values for bigrams. Like Zhu et al. they show a high correlation between values obtained from the web and values from a given (traditional) corpus (the BNC). The main application Keller et al. suggests is for bigrams that are missing in a given corpus. Nakov et al. [14] show that the n-gram count from several Internet search engines differs but is not statistically significantly different. They also show that the results from one search engine are stable over time which is encouraging for researchers using this technique to obtain $TC$ values.

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 conventional corpora. 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.

2.2 Generating IDF Values for Web Pages

Sugiyama et al. [5] use the TREC-9 Web Track dataset [15] 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 the generation of 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. They claim that this method represents the web page’s content more accurately and hence improves the retrieval performance.

Phelps and Wilensky [1] 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 the 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 Setup

The WaC corpus provides what they call a frequency list, a list of all unique terms in the corpus (lemmatized and non-lemmatized) and their $TC$ value. Since the document boundaries in the corpus are obvious, 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 dismiss all lemmatized terms and only use the non-lemmatized terms. We rank both lists in decreasing order of their TC and DF values and honor ties with the minimum value. For example consider four terms a, b, c, d where term a has the highest value, terms b and c have the same second highest value and term d has the lowest value. The ranking here would be a=1, b=2, c=2, d=4. This kind of ranking is also known as the sports ranking. We compute the Spearman’s ρ and Kendall τ to mathematically prove the correlation. Table 3 shows the top 20 terms from the WaC corpus ordered by decreasing TC and DF values. The similarity between the two rankings already becomes visible with that small example since the intersection of both top 20 rankings just holds 22 unique terms. It is not surprising that the TC values are much greater than the DF values since for DF duplicates within one document are not counted. Since Table 3 mainly holds stop words we show terms ranked between 101 and 120 and their TC and DF values in Table 4. The correlation is obviously less strong and the number of intersecting terms went up to 33.

### 4 Experiment Results

#### 4.1 Correlation within the WaC Corpus

Figure 1 shows (in loglog scale) the correlation of ranked terms within the WaC corpus. The x-axis represents the TC ranks of terms and the y-axis the corresponding DF rank of the same term. As expected we see the majority of the points within a diagonal corridor which indicates a great similarity between the rankings.

Figures 2 and 3 show the measured and estimated correlation between TC and DF values in the WaC dataset. The solid black line displays Spearman’s ρ. The increasing size of the dataset is shown on the x-axis. The value for ρ at any size of the dataset is beyond 0.8 which indicates a very strong correlation between the rankings. The results are statistically significant with a p-value of $2.2 \times 10^{-16}$. The blue line in both Figures shows the computed Kendall τ values for the top 1,000,000 ranks and the dotted red line

| Rank | Term | TC     | Term | DF     |
|------|------|--------|------|--------|
| 1    | the  | 116448129 | the  | 2662742 |
| 2    | of   | 59869301  | and  | 2635683 |
| 3    | and  | 58521777  | to   | 2620998 |
| 4    | to   | 53923142  | of   | 2613789 |
| 5    | a    | 40940103  | a    | 2582936 |
| 6    | in   | 36463498  | in   | 2533431 |
| 7    | is   | 22389310  | for  | 2321630 |
| 8    | for  | 21754176  | is   | 231630  |
| 9    | that | 16665399  | on   | 2261602 |
| 10   | on   | 15636014  | with | 2221913 |
| 11   | with | 13985141  | are  | 1996578 |
| 12   | it   | 13518855  | this | 1981575 |
| 13   | be   | 13007008  | from | 1964174 |
| 14   | as   | 11943257  | be   | 1951862 |
| 15   | are  | 11571176  | by   | 1947784 |
| 16   | you  | 11298405  | as   | 1947590 |
| 17   | this | 11218852  | that | 1943238 |
| 18   | by   | 10639772  | at   | 1927033 |
| 19   | at   | 9907466   | it   | 1857296 |
| 20   | i    | 9855628   | an   | 1788905 |
Table 4: Terms ranked between 101 and 120 and their TC and DF Values

| Rank | Term | TC  | Term | DF   |
|------|------|-----|------|------|
| 101  | ...  | 1667105 | he   | 695825 |
| 102  | get  | 1639959 | get  | 689949 |
| 103  | good | 1623519 | part | 686865 |
| 104  | her  | 1594657 | need | 684872 |
| 105  | me   | 1578177 | his  | 683221 |
| 106  | back | 1547902 | could| 680212 |
| 107  | uk   | 1538433 | those| 678384 |
| 108  | made | 1524567 | before| 671783 |
| 109  | way  | 1498196 | between| 67402 |
| 110  | need | 1498142 | here | 667614 |
| 111  | those| 1489151 | available| 660078 |
| 112  | between | 1484492 | each | 659960 |
| 113  | she  | 1482487 | n’t  | 655339 |
| 114  | 2    | 1465266 | back | 644124 |
| 115  | 1    | 1462262 | much | 638500 |
| 116  | day  | 1451911 | used | 634636 |
| 117  | service | 1439068 | including | 633138 |
| 118  | world | 1437539 | help | 632119 |
| 119  | here | 1436429 | number | 616825 |
| 120  | used | 1429151 | own  | 614981 |

Figure 1: Correlation between Term Count and Document Frequency in the WaC dataset
Figure 2: Measured and Estimated Correlation between Term Count and Document Frequency in the WaC dataset - Normal Scale

Figure 3: Measured and Estimated Correlation between Term Count and Document Frequency in the WaC dataset - Semi-Log Scale
represents the estimated values for the remaining set of data in the WaC corpus. Since the computed $\tau$ values are hard to read on a normal scale (Figure 2) we plotted the same graph in semi-log scale in Figure 3. The computed $\tau$ values vary between 0.82 and 0.74 and the estimated values have a minimum of 0.66.

We did not compute $\tau$ for greater ranks since it is a very time consuming operation and the estimated values also indicate a strong correlation. Gilpin [16] provides a table for converting $\tau$ into $\rho$ values. We use this data to estimate our $\tau$ values. Even though the data in [16] is based on $\tau$ 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. For example, it shows that a $\tau$ value of 0.8 can be converted to a $\rho$ of 0.94 which is consistent with our measured values shown in Figure 2. Therefore we can predict the high $\tau$ values even beyond the top 1,000,000 ranks shown in Figure 3.

4.2 Computation Time for Kendall $\tau$

Figure 4 shows the measured and predicted computation time (y-axis, in seconds) for $\tau$ of top $n$ rankings (x-axis). The black solid line shows the measured time values for rankings up to the top 1,000,000 terms. The red dashed line represents the predicted time values for the entire corpus and (in the small plot in the left top corner) for the top 1,000,000 ranks. Figure 4 shows the observed complexity of $O(n^2)$. For the entire WaC dataset (over 11 million unique terms) we estimate a computation time for Kendall $\tau$ of almost 11 million seconds or more than 126 days which is clearly beyond a reasonable computation time for a correlation value. Kendall $\tau$ was computed using an off-the-shelf correlation function as part of the R-Project, an open source environment for statistical computing. The software (version 2.6) was run on a Dell Server with a Pentium P4 2.8Ghz CPU and 1 GB of memory.

---

1http://www.r-project.org/
4.3 Term Count - Document Frequency Ratio in the WaC Corpus

Another interesting way to show the correlation between TC and DF values is simply looking at the ratio of the two values. Figure 5(a) shows the distribution of TC/DF ratios with values rounded after the second decimal and Figure 5(b) shows the ratios rounded after the first decimal. It becomes obvious that the vast majority of the ratio values are very small. The visual impression is supported by the computed mean value of 1.23 with a standard deviation of $\sigma = 1.21$ for both, Figure 5(a) and 5(b). The median of ratios is 1.00 and 1.0 respectively. Figure 5(c) shows the distribution of TC/DF ratios rounded as integer values. It is consistent with the pattern of Figures 5(a) and 5(b) and the mean value is equally low at 1.23 ($\sigma = 1.22$). The median here is also 1. Figure 5 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.

4.4 Correlation between the WaC and the N-gram Corpus

The TC values for both corpora, WaC and N-gram, are available and therefore we investigate their correlation. Figure 6 displays (in loglog scale) the frequencies of unique TC values in both corpora. The graph shows the TC threshold of 200 Google applied while creating the N-gram. By visual observation it becomes obvious that the distribution of TC values in both corpora is very similar. Just the size of the Google N-gram corpus is responsible for the offset between the graphs.

5 Conclusion

We have shown a very strong correlation between the TC and DF values within the WaC corpus with Spearman’s $\rho \geq 0.8$ ($p \leq 2.2 \times 10^{-16}$). This result leads us to the conclusion that the two values can be used interchangeably and therefore TC values are usable for the generation of accurate IDF values. We also show (by visual observation) a high correlation between the TC values of the WaC and of the N-gram datasets. We can now claim that, despite the fact that the Google N-gram dataset does not contain DF values, the corpus and its TC values are also usable for accurate IDF computation which can lead to the generation of LSs of web pages.

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.
Figure 5: Frequency of $TC/DF$ Ratios in the WaC Corpus
Figure 6: Term Count Frequencies in the WaC and N-gram Corpus
References

[1] Thomas A. Phelps and Robert Wilensky. Robust Hyperlinks Cost Just Five Words Each. Technical Report UCB/CSD-00-1091, University of California at Berkeley, Berkeley, CA, USA, 2000.

[2] Seung-Taek Park, David M. Pennock, C. Lee Giles, and Robert Krovetz. Analysis of Lexical Signatures for Improving Information Persistence on the World Wide Web. *ACM Transactions on Information Systems*, 22(4):540–572, 2004.

[3] Terry L. Harrison and Michael L. Nelson. Just-in-Time Recovery of Missing Web Pages. In *Proceedings of HYPERTEXT ’06*, pages 145–156, 2006.

[4] Martin Klein and Michael L. Nelson. Revisiting Lexical Signatures to (Re-)Discover Web Pages. In *Proceedings of ECDL ’08*, 2008.

[5] Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, and Shunsuke Uemura. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. In *Proceedings of HYPERTEXT ’03*, pages 198–207, 2003.

[6] Ian Soboroff. Do TREC Web Collections Look Like the Web? *SIGIR Forum*, 36(2):23–31, 2002.

[7] Geoffrey Leech, Lawrence P. Grayson, and Andrew Wilson. Word Frequencies in Written and Spoken English: based on the British National Corpus. Longman, London, 2001.

[8] Jessica Staddon, Philippe Golle, and Bryce Zimny. Web based inference detection. In *USENIX Security Symposium*, 2007.

[9] Alex Franz and Thorsten Brants. All Our N-Gram are Belong to You. [http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html](http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html).

[10] Aleksander Kolcz, Abdur Chowdhury, and Joshua Alspector. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization. In *KDD*, pages 605–610, 2004.

[11] Wei-Tsen Milly Chiang, Markus Hagenbuchner, and Ah Chung Tsoi. The WT10G Dataset and the Evolution of the Web. In *Proceedings of WWW ’05*, pages 938–939, 2005.

[12] Xiaojin Zhu and Ronald Rosenfeld. Improving Trigram Language Modeling with the World Wide Web. In *Proceedings of ICASSP ’01*, pages 533–536, 2001.

[13] Frank Keller and Mirella Lapata. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29(3):459–484, 2003.

[14] Preslav Nakov and Marti Hearst. A Study of Using Search Engine Page Hits as a Proxy for n-gram Frequencies. In *Proceedings of RANLP ’05*, 2005.

[15] David Hawking. Overview of the TREC-9 Web Track. In *NIST Special Publication 500-249: TREC-9*, pages 87–102, 2001.

[16] Andrew 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.