An Introduction to Local Weight Models

Abstract –This is Part 4 of a tutorial series on Term Vector Theory. An introduction to several local weight models is presented.

Keywords: term frequency, local weights, scaling, logs, power models, weighting schemes
Published: 10-27-2006; Updated: 06-11-2020
© E. Garcia, PhD; admin@minerazzi.com

Note: This article is part of a legacy series that the author published circa 2006 at http://www.miislita.com, now a search engine site. It is now republished in pdf format here at http://www.minerazzi.com, with its content edited and updated. The original articles can be found referenced in online research publications on IR and elsewhere.

Introduction

In Parts 1, 2, and 3 of this series on vector space models for Information Retrieval (IR) we described several models based on local, $L_{i,j}$, and global, G_i, weights. $L_{i,j}$ was defined in terms of local frequencies, $f_{i,j}$, and G_i using inverse document frequencies ($IDF_i = log(D/d_i)$) where d_i is the number of documents that mention term i in a collection consisting of D documents.

Unfortunately, these models are vulnerable to a large set of deceiving practices known as *spamdexing* (Garcia, 2016a; 2016b; 2016c; AIRWeb, 2007). Figure 1 depicts two of these models, the Binary (BNRY) and Term Count (FREQ) models. Both models have specific strengths and drawbacks.

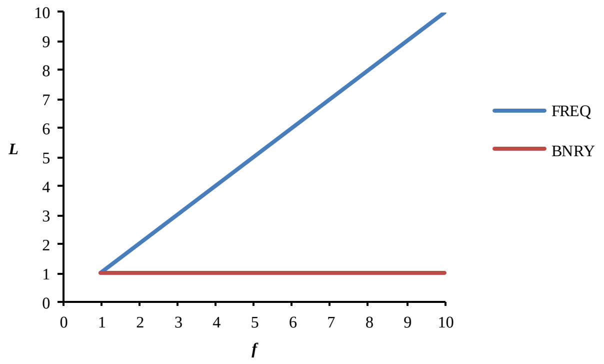

Figure 1. Binary (BNRY) and Term Count (FREQ) Models.

BNRY, $L_{i,j} \begin{cases} 1\ if\ f_{i,j} > 0 \\ 0\ if\ f_{i,j} = 0 \end{cases}$, cannot be gamed by repeating terms. However, by creating vocabulary-rich documents, the retrieval possibilities of said documents increase. By contrast FREQ, $L_{i,j} \begin{cases} f_{i,j}\ if\ f_{i,j} > 0 \\ 0\ if\ f_{i,j} = 0 \end{cases}$, is term frequency dependent and favors longer documents as these tend to repeat terms. Because FREQ assumes that terms repeated x times are x times more important, the model can be easily gamed by simply repeating terms.

Ideally, a model should be middle ground between BNRY and FREQ. What we mean by this is, not that said model should describe a curve half-way the curves shown in Figure 1, but that it should incorporate the strengths of the two models. Such a model should describe term weights that saturate after a few occurrences, being robust against term repetition.

One can find in the information retrieval literature about a dozen of $L_{i,j}$ models (Chisholm & Kolda, 1999; Lee, Chuang, Seamons,1997). Some of these are based on one or more of the following transformations:

- scaling
- logs
- powers

The purpose of this tutorial is to introduce readers to these transformations. In the next section we discuss several weighting schemes based on some of these.

Scaling-based Transformations

The simplest of these transformations consists in rescaling the range of term frequencies present in a document j to the [0,1] range with the general formula

$$L_{i,j} \begin{cases} \frac{f_{i,j} - min\, f_{i,j}}{max f_{i,j} - min\, f_{i,j}}\ if\ f_{i,j} > 0 \\ 0\ if\ f_{i,j} = 0 \end{cases} \tag{1}$$

where *minf*$_{i,j}$ and *maxf*$_{i,j}$ are the minimum and maximum term frequencies in document j.

However, (1) is not useful for weighting terms with a homogeneous presence in a document because for such documents $f_{i,j} = minf_{i,j} = maxf_{i,j}$. Dropping $minf_{i,j}$ solves this drawback

$$L_{i,j} \begin{cases} \frac{f_{i,j}}{maxf_{i,j}} \; if \; f_{i,j} > 0 \\ 0 \; if \; f_{i,j} = 0 \end{cases} \tag{2}$$

where the scale of term weights is still upper bounded.

A commonly used transformation consists in using the average term frequency of document j as the scaling factor, with the new scale no longer upper bounded; i.e.

$$L_{i,j} \begin{cases} \frac{f_{i,j}}{avef_{i,j}} \; if \; f_{i,j} > 0 \\ 0 \; if \; f_{i,j} = 0 \end{cases} \tag{3}$$

To distinguish (2) and (3) from FREQ, these can be called FREQX and FREQA, respectively. Another popular transformation consists in using the maximum term frequency of document j as the scaling factor and then applying an augmentation factor K , usually set to $K = 0.5$, to scaled frequencies, like this

$$L_{i,j} \begin{cases} K + (1 - K) \frac{f_{i,j}}{maxf_{i,j}} \; if \; f_{i,j} > 0 \\ 0 \; if \; f_{i,j} = 0 \end{cases} \tag{4}$$

For $K = 0.5$, (4) is called the augmented term frequency model (ATF1), awarding a term for its presence and repetition in a document. The scale of weights for terms present in a document is compressed to the [0.5, 1] interval. Several weighting schemes can be derived from (4).

For instance, for $K = 0$ (4) reduces to (3) while for $K = 1$ to BNRY so ATF1 is middle ground between these models. For $K < 0.5$, i.e., $K = 0.2$, (4) reduces to a new model called the ATF Changed-coefficient model (ATFC). This model awards more weight for the repetition of a term in a document and less for its mere presence. The reverse occurs for $K > 0.5$. For instance for $K = 0.90$ a new model, the augmented average term frequency model (ATFA), is obtained.

Log-based Transformations

The Binary model cannot discriminate between terms that appear once in a document and those that appear frequently in the same document. By contrast, frequency-based models tend to give too much weight to index terms that appear frequently in a document. Logarithms provide a middle ground.

Before proceeding any further, a word of caution is necessary. Do not assume that a "log" notation found in the IR literature always means decimal logs. For instance, in footnote 2 of their 1999 report, Chisholm & Kolda clarified that they used "log" to mean $\log_2$: "All logs are base two". We adopt the same convention in this tutorial.

Certainly the base of the logarithms does not matter and one could safely use logs at other bases like decimal logs, $\log_{10}$, or base e logs (natural logs, ln). To convert logs across bases, use $log_{newbase} = \frac{\ln(number)}{\ln(newbase)}$. For instance, the log of 3 in base 2 is ln(3)/ln(2) = 1.5849… ≈ 1.58.

Another observation that is worth to point out is that log transformations can produce zero and negative values. This can be offset in two different ways, leading to two different weighting schemes: by either adding 1 to the values to be transformed and then taking logs or by adding 1 after taking logs. With this in mind, the simplest model that uses log transformations is the Log model (LOGA),

$$L_{i,j} \begin{cases} 1 + \log(f_{i,j})\ if\ f_{i,j} > 0 \\ 0\ if\ f_{i,j} = 0 \end{cases} \tag{5}$$

where the "A" indicates that the log is augmented by adding 1 for the reasons explained before. In this case the scale of weights for terms present in a document is not upper bounded. To insure that the scale is upper bounded, (5) is normalized with the document average term frequency,

$$L_{i,j} \begin{cases} \frac{1+\log(f_{i,j})}{1+\log(ave\,f_{i,j})}\ if\ f_{i,j} > 0 \\ 0\ if\ f_{i,j} = 0 \end{cases} \tag{6}$$

Chisholm and Kolda (1999) call (6) the Log Normalized model (LOGN).

When no global weights are used to rank documents, it is recommended to use normalized local weights. LOGN conforms to this requirement. A combination of (4) and (5) leads to a new weighting scheme called the Augmented Log model (LOGG),

$$L_{i,j} \begin{cases} K + (1 - K)\log\left(f_{i,j} + 1\right) \; if \; f_{i,j} > 0 \\ 0 \; if \; f_{i,j} = 0 \end{cases} \tag{7}$$

where its upper bound is usually set to K = 0.2. For K = 0, (7) reduces to ALOG, Log Attenuated, $\log\left(f_{i,j} + 1\right)$, and for K = 1, to BNRY. Another log-based weighting scheme, first proposed by Harman (1986), is

$$L_{i,j} \begin{cases} \frac{\log\left(f_{i,j}+1\right)}{\log\left(length_j\right)} \; if \; f_{i,j} > 0 \\ 0 \; if \; f_{i,j} = 0 \end{cases} \tag{8}$$

where $length_j$ is the length of document j computed as the number of unique terms. This model, which we might refer to as the Log Length Normalized model (LOGLN), has been used by others (Frakes & Baeza-Yates, 1992; Crestani, Ruthven, Sanderson, & Rijsbergen, 1995; Sanderson & Ruthven, 1996) as part of the TF-IDF model

$$w_{i,j} = \frac{\log\left(f_{i,j}+1\right)}{\log\left(length_j\right)} IDF_i = \frac{\log\left(f_{i,j}+1\right)}{\log\left(length_j\right)} \log\left(\frac{D}{d_i}\right) \tag{9}$$

where documents are ranked by summing the $w_{i,j}$ weights of query terms found in the documents.

Power-based Transformations

The Square Root model (SQRT) is an example of a power transformation

$$L_{i,j} = 1 + \left(f_{i,j} - 0.5\right)^{1/2} \tag{10}$$

It was developed by realizing that $f_{i,j}$ ½ describes a curve close to that of LOGA, a top performer. This model can be derived using variance-stabilizing transformations.

Comparative

In Figures 2 and 3 we compare all of these models. The absolute positions of the *L*-*f* curves are not important; what is important is the relative increments of *L* for different increments in *f*.

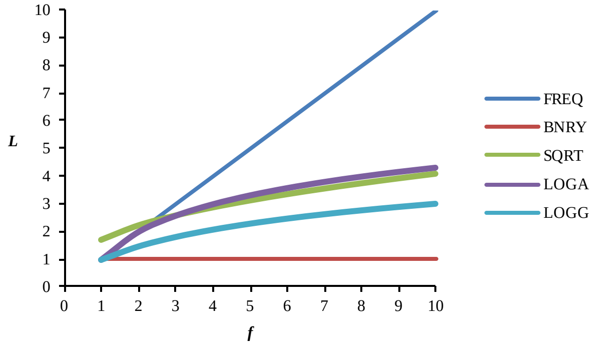

Figure 2. SQRT, LOGA, and LOGG as middle ground between FREQ and BNRY.

Notice from Figure 2 that the models are middle ground between FREQ and BNRY. SQRT returns weights similar to those of LOGA and a curve similar to that of LOGG. Eventually these curves reach a plateau corresponding to the saturation of term weights.

In Figure 3, we compared the LOGN, ATFA, ATF1, ATFC, and LOGLN models by assuming a hypothetical document *j* with *n* unique terms. We assumed that these were listed in increasing order of term frequencies, with the frequency of the last *m* terms incremented by 1, with no ties, and from *f = 1* to *f* = *m*. This implies that *f* = 1 for the first *n* – *m* terms. Therefore, $length_j = n$, $maxf_{i,j} = m$, and $avef_{i,j} = \frac{n-m+\frac{m(m+1)}{2}}{n}$.

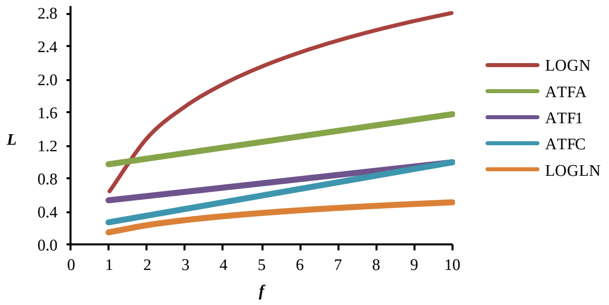

Figure 3. Comparative between the ATF1, ATFC, LOGN, ATFA, and LOGLN models.

Figure 3 corresponds to the case where n = 100 and m = 10, hence $avef_{i,j}$ = 1.45. LOGN and LOGLN describe curves that effectively saturate term weights, but LOGLN does it faster and after a few occurrences. As expected, ATF1, ATFC, and ATFA describe straight line curves. Notice that with ATF1 and ATFC, $L_{i,j}$ = 1 when $f_{i,j}$ = $maxf_{i,j}$. By contrast with ATFA, $L_{i,j}$ = 1 when $f_{i,j}$ = $avef_{i,j}$.

With LOGLN, $L_{i,j}$ = 1 if and only if f_{ij} = $length_{\,j}$ – 1. This will be the case of, for instance, a document with 100 unique terms with one repeated 99 times. An extreme, though not impossible, scenario of term repetition abuses: keyword spam.

Conclusion

Frequency-based local weights are vulnerable to keyword repetition. One way to fight against this adversarial IR practice consists in transforming raw frequencies.

We have briefly described several models that attempt to do that. Additional IR weighting schemes have been proposed (Chisholm & Kolda, 1999; Lee, Chuang, Seamons,1997), and new ones can certainly be proposed.

Best matching algorithms like BM25 (Robertson, 2004; Wikipedia, 2016) have been proposed as alternatives to the above models (Robertson, 2004, Wikipedia, 2016). Unfortunately, BM25 models are in practice difficult to implement efficiently, requiring of parameterized functions.

In recent years, binned or document-centric impact models have been developed to overcome some of these efficiency issues (Anh & Moffat, 2004; 2005; Metzler, Strohman, & Croft, 2008). BM25 as these models deserve separate tutorials.

Exercises

1. Local weight models are functions of the form $L_{i,j}(f_{i,j})$. For each of the weighting functions discussed in this tutorial, compute the $dL_{i,j}/d\,f_{i,j}$ derivative.
2. In the previous exercise, what you might conclude by comparing $dL_{i,j}/d\,f_{i,j}$ derivatives for the models discussed in this tutorial?
3. Rework the exercise given in Part 3 of this series, this time using expression (8).

References

AIRWeb (2007). Adversarial Information Retrieval on the Web. Retrieved from http://airweb.cse.lehigh.edu/2007/cfp.html

Anh, V.N. and Moffat, A. (2004). Collection-independent document-centric impacts. In: Proc. Australian Document Computing Symposium. 25 – 32. Retrieved from https://pdfs.semanticscholar.org/1779/d097851fdbbe1b3c5fe182ddab43bc4c136c.pdf.

Anh, V.N. and Moffat, A. (2005). Simplified similarity scoring using term ranks. In: Proc. 28th SIGIR. 226 – 233. Retrieved from http://www2.dcc.ufmg.br/eventos/sigir2005/files/talks-papers-2005-08-11/AlistairMoffat.pdf

Chisholm, E. and Kolda, T. G. (1999). New Term Weighting Formulas for the Vector Space Method in Information Retrieval. Oak Ridge National Laboratory. Retrieved from http://www.sandia.gov/~tgkolda/pubs/pubfiles/ornl-tm-13756.pdf

Crestani, F., Sanderson, M., Ruthven, M. I., and Rijsbergen, C. J. (1995). The troubles with using a logical model of IR on a large collection of documents. Proceedings of the 4th TREC conference (TREC-4). NIST, Pages 509-526. Retrieved from http://marksanderson.org/publications/my_papers/TREC-4-Notebook.pdf

Frakes, W. B. and Baeza-Yates, R. (1992). *Information Retrieval: Data structures and algorithms.* Chapter 14. Prentice Hall. Retrieved from http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap14.htm

Garcia, E. (2016a). Term Vector Theory and Keyword Weights. Retrieved from http://www.minerazzi.com/tutorials/term-vector-1.pdf

Garcia, E. (2016b). The Binary and Term Count Models. Retrieved from http://www.minerazzi.com/tutorials/term-vector-2.pdf

Garcia, E. (2016c). The Classic TF-IDF Vector Space Model. Retrieved from http://www.minerazzi.com/tutorials/term-vector-3.pdf

Harman, D. (1986). An Experimental Study of Factors Important in Document Ranking. Paper presented at ACM Conference on Research and Development in Information Retrieval, Pisa, Italy. Retrieved from https://www.researchgate.net/publication/221301152_An_Experimental_Study_of_Factors_Important_in_Document_Ranking

Lee, D. L., Chuang, H., and Seamons (1997). Document Ranking and the Vector-Space Model. IEEE March/April, pp 67-75. Retrieved from http://www.cs.ust.hk/faculty/dlee/Papers/ir/ieee-sw-rank.pdf

Metzler, D., Strohman, T., and Croft, W. B. (2008). A Statistical View of Binned Retrieval Models. Advances in Information Retrieval. Volume 4956 of the series Lecture Notes in Computer Science pp 175-186. Springer. Retrieved from https://pdfs.semanticscholar.org/0814/4a3c963e7af72c0756abf01fbc2551ed7f89.pdf

Robertson, S. E. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation, 60, 5, 503-520. Retrieved from http://nlp.cs.swarthmore.edu/~richardw/papers/robertson2004-understanding.pdf

Sanderson, M. and Ruthven, I. (1996). Report on the Glasgow IR group (glair4) submission. Proceedings of the 5th TREC conference (TREC-5). NIST, Pages 517-520. Retrieved from http://marksanderson.org/publications/my_papers/TREC-5_report.pdf

Wikipedia (2016). Okapi BM25. Retrieved from https://en.wikipedia.org/wiki/Okapi_BM25