Theoretical Bounds On and Empirical Robustness of Score Regularization to Different Similarity Measures

Fernando Diaz[*]
Yahoo! Inc.
1000 Rue de la Gauchetiere, Suite 2400
Montreal, QC
diazf@yahoo-inc.com

ABSTRACT

We present theoretical bounds and empirical robustness of score regularization given changes in the similarity measure.

Categories and Subject Descriptors: I.5.3 Clustering: Similarity Measures

General Terms: Reliability

Keywords: regularization, inter-document similarity

1. INTRODUCTION

Score regularization refers to the process of re-ranking documents so that topically related documents receive more consistent scores [2]. Regularization has been demonstrated to significantly improve retrieval performance. A fundamental data structure in the regularization algorithm is the inter-document affinity matrix. This inter-document similarity matrix can be computed using any number of term-based similarity measures [1]. According to van Rijsbergen, text affinity measures share heuristics which result in very similar behavior [4, page 24]. In this poster, we will establish theoretical bounds and present empirical evidence of the effect different similarity measures have on regularization.

We will view regularization as the solution of a linear system. Given the original scores $\mathbf{y}$, we rewrite the closed form version of regularization ([2, Equation 15]), $\mathbf{f}^*$, as,

$$\left(\frac{\alpha}{1-\alpha}\Delta + \mathbf{I}\right)\mathbf{f}^* = \mathbf{y} \tag{1}$$

where the Laplacian, Δ, is associated with the matrix, $\mathbf{W}$, generated by some arbitrary similarity measure (for example, cosine similarity) and α controls the amount of regularization.

2. THEORETICAL RESULTS

We consider a matrix, $\tilde{\mathbf{W}}$, generated by a different similarity matrix (for example Hellinger similarity). The regularized scores using this alternative similarity measure, $\tilde{\mathbf{f}}$, is the solution to the linear system in Equation 1 using the Laplacian, $\tilde{\Delta}$, associated with $\tilde{\mathbf{W}}$.

We would like to bound the difference in regularized scores given differences in the similarity matrix. We will measure the change in regularized scores using the relative error between scores,

$$\frac{\|\tilde{\mathbf{f}}^* - \mathbf{f}^*\|_2}{\|\mathbf{f}^*\|_2} \tag{2}$$

We will measure the difference in the similarity matrix according to the changes in the associated Laplacians,

$$\left\|\tilde{\Delta} - \Delta\right\|_2 \tag{3}$$

where the matrix norm is induced from the vector 2-norm.

Theorem 2.1.

$$\frac{\|\tilde{\mathbf{f}}^* - \mathbf{f}^*\|}{\|\mathbf{f}^*\|} \leq \frac{\alpha}{1-\alpha}\left\|\tilde{\Delta} - \Delta\right\|$$

We omit the proof due to space constraints.

Noting that $\|\tilde{\Delta} - \Delta\| \leq 2$ (proof omitted due to space constraints), we depict the bound on $\frac{\|\tilde{\mathbf{f}}^* - \mathbf{f}^*\|_2}{\|\mathbf{f}^*\|_2}$ in Figure 1. The general behavior of this bound confirms an intuition we might have already. For low values of α, when two affinity measures are very similar, their regularized scores are very similar. In fact, for low values of α, the regularization is quite robust to arbitrary differences in the affinity. However, as we regularize more aggressively by using a higher α, the regularized solutions are more sensitive to perturbations of the affinity matrix.

3. EMPIRICAL ROBUSTNESS OF REGULARIZATION

The range of differences, $0 \leq \|\tilde{\Delta} - \Delta\| \leq 2$, includes arbitrary matrix perturbations. In reality, our perturbations are likely to be constrained to differences much less than the maximum. In order to measure the empirical perturbations, we computed the differences between Laplacians using cosine similarity of tf.idf vectors and the Hellinger similarity of language models. Our initial scores, $\mathbf{y}$, were query likelihood retrievals of TREC queries 51-200 associated with the news portions of Tipster disks 1 and 2 [5]. We found that the mean value of $\|\tilde{\Delta} - \Delta\|$ was 0.541 ± 0.0585; this range is depicted in Figure 1. An expected perturbation in this range indicates that regularized scores will be very similar for $0 \leq \alpha \leq 0.5$. We plot the empirical regularization differences for various α for a fixed query with $\|\tilde{\Delta} - \Delta\| = 0.510$ in Figure 2. From

[*] Work conducted at the Center for Intelligent Information Retrieval.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

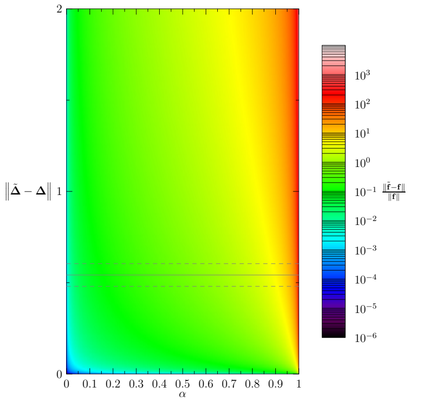

Figure 1: Bound on regularization error given similarity matrix perturbations and α. The solid horizontal line represents the empirical mean perturbation found in our experiments. The dashed lines represent one standard deviation. This figure is ideally viewed in color.

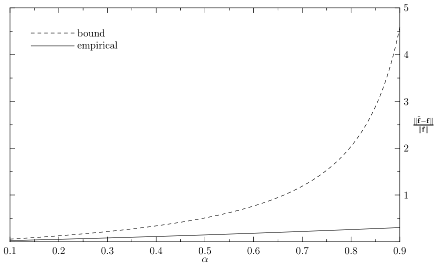

Figure 2: Empirical differences in regularized scores as a function of α for a retrieval from our experiments. This dashed line in this graph represents the theoretical bound and therefore is a cross-section of surface from Figure 1.

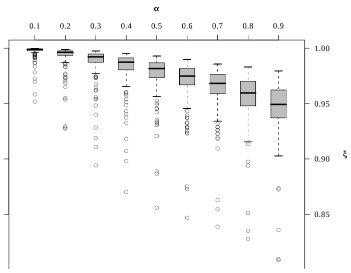

Figure 3: Empirical relationship between α and the Plantagenet coefficient.

this figure, it should be clear that our bound, because it considers arbitrary perturbations, is somewhat loose. The empirical evidence from other queries confirms that the difference between these two affinity measures is likely to be far below our bound.

The bound established in Theorem 2.1 measures the effect of different similarity measures on norm of the difference between the regularized scores. Because information retrieval is often evaluated by the induced ranking, it is worth exploring the effect on rankings resulting from different similarity measures. Therefore, for each pair of regularized rankings in our experiment, we compute the Plantagenet coefficient of rank similarity [3]. The Plantagenet is defined as,

$$\xi_n = -\frac{4n+5}{n-1} + \frac{6}{n^3-n}\sum_{i=1}^{n} x_i y_i \left(4 - \frac{x_i + y_i}{n+1}\right) \qquad (4)$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors containing ranks of each document. The Plantagenet coefficient is a version of Spearman correlation sensitive to changes in the top ranks. This measure is appropriate because, when comparing two rankings, we are most concerned with changes of the ranks of top-ranked documents. In Figure 3, we plot this correlation as a function of α. This figure gives us an intuition for the perceptible changes resulting from using different similarity measures. In fact, we see that, for all values of α, we achieve strong correlation between rankings.

In summary, we have studied the stability of regularization subject to changes of the parameter α. We found that, for small values of α, solutions are robust to small perturbations in the similarity matrix. For large α, regularization is more aggressive and solutions are more sensitive to changes in similarity. We complemented these theoretical results with empirical measurements of the effect of similarity matrix perturbations. We found that the differences between vector space model and language model similarities resulted only in slight differences in regularized scores and rank ordering of regularized scores. In future work, we would like to examine the empirical difference in regularization performance for other similarity measures.

4. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the author and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] F. R. Chen, A. O. Farahat, and T. Brant. Multiple similarity measures and source-pair information in story link detection. In *HLT/NAACL 2004*, pages 313–320, May 2004.

[2] F. Diaz. Regularizing query-based retrieval scores. *Information Retrieval*, 10(6):531–562, December 2007.

[3] C. Genest and J.-F. Plante. On Blest's measure of rank correlation. *Canadian Journal of Statistics*, 31(1):1–18, 2003.

[4] C. J. van Rijsbergen. *Information Retrieval*. 1979.

[5] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. 2001.