Length-independent vector-space document similarity measures based on regression residuals

Derrick Higgins
Educational Testing Service
dhiggins@ets.org

1 Background

One aspect of vector-space semantic similarity estimates which has so far received little attention is their dependence on the length of the texts to be compared. A simple experiment will demonstrate this effect.

The data set used for this demonstration involves texts from the Lexile collection, a set of general fiction titles spanning a wide range of grade-school reading levels, which ETS licenses from the Metametrics Corporation. 400 documents were selected randomly from this collection, and truncated so that 100 included only the first 500 word tokens, 100 included only the first 1000 words, 100 included only the first 5000 words, and the final 100 included only the first 10,000 words. Two sets of similarity scores were calculated for each pair of documents in this collection (excluding duplicates).

The first set of similarity scores was created using a simple content vector analysis (CVA) model with \( tf*idf \) weighting, with log term weights and an inverse document frequency term equal to \( \log(\frac{NDocs}{DocFreq}) \) for each term \( k \), where the document frequency estimates were derived from the TASA corpus of high-school level texts on a variety of academic subjects.

The second set of similarity scores was created using a Random Indexing (RI) (Sahlgren, 2006) model with similar parameters as those used for the CVA model. The RI model used co-occurrence of words within the same document in the TASA corpus as the basis for dimensionality reduction, and also used the TASA corpus to estimate inverse document frequency values for individual terms. Document vectors were produced as the \( tf*idf \)-weighted sum of term vectors occurring within the document, again with log weighting of both term frequencies and inverse document frequencies.

For each of these methods, cosine was used as the similarity metric.

It turns out that the similarity scores calculated by these methods for the Lexile data set are highly correlated with the variable \( gTypes \), the geometric mean of the number of word types in the two documents to be compared. The correlation between CVA similarity and \( gTypes \) is 0.86, while the correlation between RI similarity and \( \log(\text{gTypes}) \) is 0.89.

At least in part, this dependency of similarity on length is due to statistical properties of vector-space semantic techniques themselves, rather than to particular differences in text composition between short and long documents. As the documents increase in length, the law of large numbers indicates that their semantic vectors will converge to the mean vector of the distribution, and therefore that the similarity between vectors will converge to \( \text{sim}(\vec{d}_{\text{mean}}, \vec{d}_{\text{mean}}) = 1 \). Even if the documents are not drawn from the same distribution, or on similar topics, we can assume that document topics are not so sharply delimited in terms of their vocabulary that the vectors for two different documents will tend to converge to orthogonal vectors as they increase in length.

2 Previous Work

There has been earlier work in which the confounding effect of document length on vector-space methods for semantic similarity calculation has been reported (Penu-matsa et al., 2006; Higgins, 2007), but none of this previous work has taken the further step of developing an improved similarity metric.

3 Factoring out Document Length

One way in which the effect of a confound \( X \) on a variable \( Y \) can be factored out is to estimate a linear regression of \( Y \) based on \( X \), and then use the residuals of that regression equation in further analysis. In this case, we would like to remove the effect of document length on our CVA and RI similarity scores, so standard linear regression techniques were applied to estimate the slope parameter \( m \) and intercept parameter \( b \) in the equation

\[
\hat{\text{sim}}(d_1, d_2) = m \times gTypes(d_1, d_2) + b,
\]

so as to minimize the least-squares error on the Lexile data set, between the vector-space similarity scores and the estimates of these scores based on \( gTypes \). This optimization was done independently for the CVA and RI models. The residual of the regression is given by

\[
\tilde{\text{sim}}(d_1, d_2) = \text{sim}(d_1, d_2) - \hat{\text{sim}}(d_1, d_2),
\]

and this was used as a new similarity estimate in our further analyses.

To evaluate the effect of using these new residual-based similarity estimates, experiments were conducted on two publicly-available text categorization data sets. In
each of these experiments, the task was addressed using the 148 decision tree classifier from the Weka machine learning toolkit. The features used in training the classifier were the semantic similarity between the document and the set of training documents belonging to a given topic (one feature for each topic), and the number of word types found in the document to be classified (a measure of document length).

The results provided here were produced using 10-fold cross-validation within the evaluation set.

### 3.1 Data

The *TechTC* data set (Davidov et al., 2004) is actually a collection of 100 different test sets for text categorization. These documents were automatically acquired from the Web, and the topics correspond to Open Directory categories. For the experiments reported here, half of the positive exemplars and half of the negative exemplars were used as the test data for each set, and the remaining exemplars were used to produce the topic vectors to which test documents were compared. The accuracy statistics reported below reflect the average classification accuracy over all 100 topics in the TechTC set.

The *Reuters-21578* Text Categorization Collection (Lewis, 2007) is a set of Reuters newswire articles, categorized into seven topics. In the experiments described below, the “modified Apte” split (Apte et al., 1994) was used to select documents for use in training and in evaluation. Documents belonging to multiple categories were excluded from the evaluation.

### 3.2 Results

The initial results of these experiments indicated that using residuals as modified similarity scores yielded lower accuracy in almost every instance. (Compare the row “0” with the row “1” in Table 1).

However, further investigation shows that improved classification accuracy can be obtained by reducing the magnitude of the correction for the document length component. If a revised similarity estimate is calculated as

$$\text{sim}(\vec{d}_1, \vec{d}_2) - \mu \times \tilde{\text{sim}}(\vec{d}_1, \vec{d}_2),$$

where $0 < \mu < 1$, we derive a document similarity metric whose correlation with $g$Types is not zero, but is nevertheless lower than that of the original vector-space method. Table 1 demonstrates how the performance of each text classification model changes with the parameter $\mu$. For both text categorization data sets, and both types of vector-space models, classification accuracy follows the same profile. As the correction for document length is applied with a low value of $\mu$, this results in improved performance, which reaches a peak value for some $\mu$ less than 1, and then falls off as the correction approaches the least-squares optimal solution ($\mu = 1$).

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>CVA</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Reuters</em></td>
<td><em>TechTC</em></td>
</tr>
<tr>
<td>$0$</td>
<td>85.94%</td>
<td>58.40%</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>86.48%</td>
<td>58.56%</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>85.94%</td>
<td>58.81%</td>
</tr>
<tr>
<td>$\frac{3}{4}$</td>
<td>85.74%</td>
<td>58.85%</td>
</tr>
<tr>
<td>$1$</td>
<td>85.70%</td>
<td>59.58%</td>
</tr>
<tr>
<td></td>
<td>85.86%</td>
<td>59.08%</td>
</tr>
</tbody>
</table>

Table 1: Performance on text categorization tasks as a function of proportion $\mu$ of optimal regression weight.

For the TechTC data set, the difference between the best-performing value of $\mu$ and the unmodified similarity scores ($\mu = 0$) is statistically significant both for CVA ($p < .05$) and for RI ($p < .005$). For the smaller Reuters data set, the difference does not reach the .05 level of statistical significance, but its consistency with the TechTC results is nevertheless suggestive.

These findings suggest that the observed correlation of document similarity with document length may not be entirely due to statistical artifacts. It may be that some of this correlation is due to a genuine tendency for longer documents to address more similar topics, or to use more similar vocabulary, at least in the two data sets investigated here.

### References


Derrick Higgins. 2007. Sentence similarity measures for essay coherence. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7).*

