Rutgers’ HARD Track Experiences at TREC 2004


Goal

- To test techniques for using knowledge about various aspects of the information seeker’s context to improve IR system performance.
- We were particularly concerned with such knowledge which could be gained through implicit sources of evidence, rather than explicit questioning of the information seeker. We therefore used metadata rather than the clarification form.
Hypotheses

H1: Assessors with low familiarity with a topic will find that documents on topic that have high readability are relevant, and those with low readability are not relevant; assessors with high familiarity, vice versa.

H2: Assessors with low familiarity with a topic will find that documents on topic with a low number for average syllables per word are relevant, and those with a high number are not relevant; assessors with high familiarity, vice versa.

H3: Assessors with low familiarity with a topic will find that documents on topic which have a high ratio of concrete to abstract words are relevant, and those with a low ratio of concrete to abstract terms are not relevant; assessors with high familiarity, vice versa.

H4: The subjectivity or objectivity of a document will determine its membership in a genre; modifying the ranks of documents according to how well they fit to the desired genre will increase performance.

H5: Classification of documents by subjectivity or objectivity using an SVM procedure will lead to better performance than classification using a linear regression model.

H6: Documents of a single genre will have language models characteristic of that genre; the similarity, as measured by Kulbach-Leibler distance between a document’s language model and the relevant genre language model, indicates whether that document is a member of the genre; modifying the ranks of documents according to how well the fit to the desired genre will increase performance.
Hypotheses (cont.)

H7: Documents of a single genre will include terms characteristic of that genre; re-ranking lists of documents retrieved according to a topically-based query by running a query containing both the topical terms and the relevant genre terms will increase performance.

H8: The vocabulary of documents is specific for the geographic area that they refer to. More specifically, US documents can be distinguished from non-US documents.

H9: Names of US states and important US cities are an indication of US documents; names of other countries are an indication of non-US documents.

Improvement methods

For Genre:
- Regression model
- KL distance
- Query expansion
- Language models

For Familiarity:
- Flesch scores
- Syllables / word
- Abstract / concrete words

For Geography:
- Language models
Combination of Evidence

Simple weighted average
- The new scores are the weighted average between the old score and various sets of metadata scores.

Z-scores based combination
- The assumption is that the metadata scores are normally distributed and that values within one standard deviation from the mean should have little or no influence on the baseline.

\[ \text{new} \_\text{score} = \text{old} \_\text{score} \pm d \* z \* k, \]

\[ d = \frac{\text{max} \_\text{score} - \text{min} \_\text{score}}{n - 1}, \quad z = \frac{\text{metadata} \_\text{score} - \mu}{\sigma}, \]

\[ k = \text{confidence} \_\text{level} \]
## Evaluation of two runs

<table>
<thead>
<tr>
<th></th>
<th>HARD</th>
<th>SOFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(SD)</td>
<td>0.1697 (0.1999)</td>
<td>2.356 (2.789)</td>
</tr>
<tr>
<td>Run 1 (SD)</td>
<td>0.1845** (0.2065)</td>
<td>2.644* (3.053)</td>
</tr>
<tr>
<td>Run 2 (SD)</td>
<td>0.1624 (0.1983)</td>
<td>2.333 (2.852)</td>
</tr>
</tbody>
</table>

## Conclusions

1) Use iterative experimental design: test the most important hypotheses first, then extend the experiment.

2) Prefer approaches independent of training set: readability scores, concreteness/abstractness, style features, etc, rather than models of the exemplars in the training set.

**Successful:**
- Document concreteness and document readability for users with low familiarity to the topic.
- Geography models of non-US documents.

**Unsuccessful:**
- Genre models based on training topics.