Document Features Predicting Assessor Disagreement

Praveen Chandar
Department of Computer &
Information Sciences
University of Delaware
Delaware, USA
pcr@cis.udel.edu

William Webber
College of Information Studies
University of Maryland
Maryland, USA
wew@umd.edu

Ben Carterette
Department of Computer &
Information Sciences
University of Delaware
Delaware, USA
carteret@cis.udel.edu

ABSTRACT

The notion of relevance differs between assessors, thus giving rise to assessor disagreement. Although assessor disagreement has been frequently observed, the factors leading to disagreement are still an open problem. In this paper we study the relationship between assessor disagreement and various topic independent factors such as readability and cohesiveness. We build a logistic model using reading level and other simple document features to predict assessor disagreement and rank documents by decreasing probability of disagreement. We compare the predictive power of these document-level features with that of a meta-search feature that aggregates a document's ranking across multiple retrieval runs. Our features are shown to be on a par with the meta-search feature, without requiring a large and diverse set of retrieval runs to calculate. Surprisingly, however, we find that the reading level features are negatively correlated with disagreement, suggesting that they are detecting some other aspect of document content.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

Keywords

Retrieval experiment, evaluation

General Terms

Measurement, performance, experimentation

1. INTRODUCTION

Human assessors are used in information retrieval evaluation to judge the relevance of a document for a given topic. Assessors frequently disagree on the relevance of a document to a topic, however. A study by [7] found that the probability that a second assessor would agree with a first assessor's judgment that a document was relevant was only two in three. A survey of such studies done by [2] found similar results as well. While [7] found that assessor disagreement had limited effect on the comparative evaluation of systems, it does have a major impact upon the evaluation of their absolute effectiveness. Moreover, a simulation study by [4] suggests that the effect on comparative evaluation depends upon the nature of disagreement, and that an overly liberal (or careless) assessor introduces considerable noise even to the comparison of retrieval systems.

While assessor disagreement has been frequently observed, and its effect on retrieval evaluation somewhat studied, less work has been done on the factors that lead to assessor disagreement. [9] observes that there is great variability in disagreement between different assessor pairs and on different topics. Regarding assessor-level effects, [8] find that assessor training has little effect on reliability (legally trained assessors no more than untrained assessors on e-discovery tasks). Regarding topic-level effects, [11] find that more detailed assessor instructions do not seem to increase disagreement.

In addition to assessor-level and topic-level effects on assessor disagreement, there may be document-level effects: some documents may be more likely to provoke assessor disagreement than others. [10] have begun work in this direction, using metarank information across multiple runs to predict disagreement. If one assessor finds a document relevant, but it is generally lowly ranked by retrieval systems, then a second assessor is likely to disagree with the original assessor, and conversely with originally-irrelevant but highly-ranked documents.

In the current paper, we investigate the relation between assessor disagreement and various topic-independent document features. One set of such features are various metrics of the reading level or reading difficulty of a document. Our hypothesis is that documents that are more difficult to read will provoke higher levels of assessor disagreement. We also consider document length (hypothesizing that longer documents will provoke more disagreement) and document coherence (hypothesizing that less coherent documents will provoke more disagreement). Finally, we extend the metarank method of [10] by considering not only average rank across different retrieval systems, but also the variability in the ranking—using disagreement between retrieval systems as a predictor of disagreement between human assessors.

If reliable document-level predictors of assessor disagreement can be found, then they can be used to efficiently direct multiple assessments towards those documents most likely to provoke assessor disagreement. We consider this as a ranking problem, in which documents must be ranked by decreasing probability of assessor disagreement, examining the case in which this ranking must be made without any initial relevance assessment having been performed. Our experimental results indicate that document-level features give a significant improvement over random choice in predicting assessor disagreement. Moreover, where initial relevance

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

assessments are not available, document-level features predict assessor disagreement as strongly as meta-rank features, without requiring a large and diverse set of retrieval runs to calculate.

One surprise of the study is that while reading level features are predictive of assessor disagreement, the correlation is the opposite of that posited in our hypothesis: documents scored as easier to read are more, not less, likely to provoke assessor disagreement than those scored as difficult to read. This suggests that reading level features are themselves correlated with some other aspect of document construction or content, which if more directly identified could lead to even stronger predictors of assessor disagreement; a question which is left to future work.

The remainder of the paper is structured as follows. A description of our logistic regression model along with all the document-level features is given in Section 2. Section 3 describes our experiments along with the dataset used in this work, and a detailed analysis of our results is given in Section 4. Section 5 summarizes our findings and sketches future work.

2. ASSESSOR DISAGREEMENT

Our approach to the problem of predicting assessor disagreement consists of two main components: identifying features, and developing a modeling technique.

2.1 Logistic regression

We predict the probability that a document will attract divergent binary relevance assessments from two or more assessors (D), based upon various document level features $s = \langle s_i \rangle$, as $p(D = 1|s)$. As we are predicting a probability, it is natural to apply a logistic regression to this problem:

$$p(D=1|s) = \frac{e^{\beta_0 + \sum_i \beta_i s_i}}{1 + e^{\beta_0 + \sum_i \beta_i s_i}} \quad (1)$$

where s_i is the score for feature i, and the probability p is the predicted value. The fitted value β_0 in Equation 1 is the intercept, which gives the log-odds of disagreement when the score is 0, while β_i is the score coefficient for feature i, which gives the change or "slope" in log odds of disagreement for every one point increase in the given feature scores. The slope gives the strength of relationship between feature scores and probability of disagreement, while intercept the shifts the regression curve up or down the score axis.

A model can be built for each topic individually, or a universal model can be built using all queries in our dataset. The degree to which a universal model is a good approximation for per-topic models depends upon the strength of per-topic factors in influencing disagreement. The closer the universal model is to the per-topic models, the more likely it is that a generalized model can be built, that is able to predict assessor disagreement on new collections based only the feature scores.

2.2 Document Features

In this section, we discuss in detail the various predictors that we use in Equation 1 to estimate assessor disagreement. The logistic model described in Section 2.1 relies heavily on the feature scores and identifying good predictors of disagreement is critical. We use a combination of simple document characteristic features and reading level features to estimate disagreement.

2.2.1 Simple Document Features

The simple document quality features are described below:

- ***docLength*** Total number of terms in a document is a simple feature that estimates the amount of information available in the document.
- ***aveWordLength*** Average word length (number of characters) is a very simple estimate of readability of a document.
- ***Entropy*** An estimate of document cohesiveness can be obtained using the entropy of the document [3]. Document entropy is computed over the words in the document as follows:

$$\mathcal{E}(D) = -\sum_{w \in D} P(w) log(P(w)) \quad (2)$$

where $P(w)$ can be estimated by the ratio of frequency of the word to the total number of words in the document. Lower entropy reflects a document that is focused on a single topic, while higher entropy indicates a more diffuse document.

2.2.2 Reading Level Features

We employ a number of standard metrics of reading level, based upon simple textual statistics. More complicated statistical and language model approaches are left for future work [5].

- ***FleschIndex and Kincaid*** are designed to capture the comprehension level of a passage. The two measures use word and sentence length with different weighting factors. FleschIndex is a test of reading ease with higher scores indicating text that is easier to read. Kincaid is a grade score that is negatively correlated to FleschIndex. A generic formula for both metrics is given below:

$$a \cdot \frac{words}{sentences} + b \cdot \frac{syllables}{words} + c \quad (3)$$

where the values of a,b, and c are as follows: FleschIndex ($a = -1.01, b = -84.6, c = 206.83$) and Kincaid ($a = 0.39, b = 11.8, c = -15.59$).

- ***FogIndex*** relies on average sentence length and the percentage of complex words for each passage of 100 words. Words with three or more syllables are identified as complex words.

$$0.4\left[\left(\frac{words}{sentences}\right) + 100\frac{complexWords}{words}\right] \quad (4)$$

- ***SMOG*** (Simple Measure of Gobbledygook) was designed as an easier and more accurate substitute to FogIndex, and is more prevalent in the medical domain. It relies on two factors: the number of polysyllables (words with 3 or more syllables) and the number of sentences.

$$1.043\sqrt{numOfPolysyllables \times \frac{30}{sentences}} + 3.129 \quad (5)$$

- ***Lix*** is a simple measure of readability computed by adding average sentence length and number of long words. Words with 6 or more letters are considered as long words.

$$\frac{words}{sentences} + \frac{(longwords \times 100)}{words} \quad (6)$$

- ***ARI*** (Automated Readability Index) is computed by combining the ratio of the number of characters per word and number of words per sentence. ARI relies on the number of characters per word instead of syllables per word.

$$4.71\frac{characters}{words} + 0.5\frac{words}{sentences} - 21.43 \quad (7)$$

- ***Coleman-Liau*** is very similar to the ARI, computed by a linear combination of average number of letters per 100 words and average number of sentences per 100 words.

$$0.059L - 0.296S - 15.8 \tag{8}$$

where, L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

2.2.3 Metarank Feature

[10] propose using the metarank of a document across multiple retrieval runs as a predictor that a second assessor would disagree with an original assessor, given the original assessor's judgment. The metarank method used was the meta-AP score of [1], which is a document's implicit average precision (AP) weight in a ranking. [10] used average meta-AP score as their predictor. We add to this, maximum meta-AP score and standard deviation of meta-AP scores, the last of which is a measure of the disagreement between retrieval systems over what rank a document should be returned at. Note also that [10] assume that the assessment of the original assessor was available, and build separate models for the originally-relevant and originally-irrelevant conditions; in this paper, however, we assume no assessments have been made, and build a single model to predict assessor disagreement.

3. EXPERIMENT DESIGN

3.1 Data

We use the multiply-assessed TREC 4 AdHoc dataset described by [7]. The dataset consists of 48 topics, with up to 200 relevant and 200 irrelevant pooled documents selected for multiple assessment by two alternative assessors, additional to the assessment of the topic author (who we refer to as the original assessor). We restrict ourselves only to documents from the Associated Press subcollection to avoid biases introduced by the non-random method of selecting documents for multiple assessment, and follow [7] in dropping Topics 201 and 214, as the original assessor found no documents relevant for the former, and the first alternative assessor found none relevant for the latter. We regard the assessment of a document as "disagreed" if the three relevance assessors do not all give the same assessment; this is the condition that our model will attempt to predict.

3.2 Implementation

We build per-topic models (Section 2.1) for performing feature analysis (Section 4.1), but a universal model for ranking by predicted disagreement (Section 4.2), since we assume that it is redundant to perform multiple assessments just to train up per-topic models in practice; learning-to-rank methods that adapt models for topics is left to future work. The model construction and evaluation method used in the disagreement ranking stage is described below.

- *Normalization* - Prior research has found the range of reading level scores to vary greatly with the topic [6]. It is a reasonable approach to normalize feature scores, making scores and models more stable across topics. We use the following L1 normalization to normalize the scores of each feature for each topic.

$$norm(x) = x/sum(x) \tag{9}$$

- *Cross-Validation* - We test the generalizability of the predictive ranking method using cross-validation. The dataset of 48 topics is split into 5 folds; one fold is held out for testing, and the other four used to develop a universal model. This avoids having a topic in both training and testing sets.

- *Training* - Each query in the training data is used to build a logistic model as described in Section 2.1. The maximum likelihood approach to fit the data provides us with values of intercept β_0 and coefficients β_i in Equation 1. Finally, the intercept and coefficients of our training model are obtained by computing the mean intercept and coefficients values over all training queries.

- *Testing* - The feature scores are computed for each document in the unseen test query. The probability of disagreement is obtained using Equation 1 by substituting s_i from the computed feature scores, and intercept β_0 and coefficients β_1 from the trained model. Sorting documents by decreasing order of probability of disagreement gives the final ranked list.

We evaluate the quality of the rankings of documents by probability of disagreement using 11 point precision–recall curves, mean average precision, and precision at various cutoffs, with the ground truth being documents that the three assessors disagree upon the relevance of.

4. RESULTS AND ANALYSIS

We first analyze the relationship between individual features and assessor disagreement by performing per-topic regressions (Section 4.1), then investigate the usefulness of these features as predictors of disagreement by building and testing universal (cross-validated) models (Section 4.2).

4.1 Feature Analysis

We test our hypotheses that: (1) documents with higher comprehension difficulty, (2) longer documents, and (3) documents that are less focused on a topic (less cohesive), are more likely to be disagreed upon. For each feature, we build a logistic regression model on each topic with that feature as the single predictor, and observe the coefficients that the feature achieves across the 48 topics (the β values in Equation 1). We calculate the average coefficient, and perform a two-sided, one sample t-test to test whether this coefficient differs significantly from zero across the 48 topics.

Table 4.1 reports our results. The metarank features are all highly significant. Entropy is also a significant positive predictor. In so far as entropy measures topic diffuseness, this confirms our hypothesis that more diffuse documents provoke higher levels of disagreement. Many of the reading level predictors also prove significantly correlated with disagreement. Surprisingly, however, the correlation is in the opposite direction from the hypothesis. Documents that get lower reading level scores, and therefore are marked as being easier to reading, in fact provoke higher levels of assessor disagreement. (Recall that FleschIndex is the only reading level feature where higher scores mean easier comprehension.)

4.2 Modeling Disagreement

Next, we investigate how useful our method is at predicting assessor disagreement, using a universal (cross-validated) model to rank the documents of each topic by decreasing probability of assessor disagreement. Table 4.2 summarizes performances for average precision and precision at various cutoffs. We add as a baseline the expected precision achieved by a random sorting of the documents, which is just the macroaveraged proportion of disagreed documents per topic. A universal model that combines all our

Predictor	p-value	β_i
FleschIndex	0.108	139.4
ColemanLiau	0.163	-164.4
SMOGGrading	0.077	-166.4
Lix	0.012	-241.7
Kincaid	0.022	-133.3
ARI	0.006	-156.0
FogIndex	0.018	-159.2
docLength	0.052	51.2
aveWordLength	0.225	-374.7
Entropy	< 0.001	832.1
metaAPSum	< 0.001	159.7
metaAPStDev	< 0.001	206.8
metaAPMax	< 0.001	321.2

Table 1: Results of significance test using two-sided one sample t-test with p-values and mean co-efficient scores across all 48 topics.

Predictor	P@5	P@10	P@20	MAP
random	0.216	0.216	0.216	0.216
metaAP	0.317*	0.350*	0.357*	0.372*
docLength	0.229	0.229	0.235	0.255*
Entropy	0.258	0.254	0.241	0.261*
aveWordLength	0.200	0.190	0.215	0.240*
ReadingLevel	0.246	0.252	0.229	0.239*
All Combined	0.321*	0.329*	0.341*	0.362*

Table 2: Performance Comparison at various ranks with significant improvement over expected random scores indicated by * (paired t-test). The results are based on 5-fold cross validation across 48 topics.

features (denoted by "All Combined") and a model that uses the metarank features significantly improves over random ordering under all measures. All the other features give a significant improvement over random order for MAP only, suggesting that top-of-ranking performance is mediocre. Entropy does best, as in Table 4.1, whereas the combined reading levels, despite being significant correlated with disagreement give very little benefit in terms of predicting disagreement under a universal model.

5. CONCLUSION

We started this paper with three hypotheses, namely that the documents that assessors are more likely to disagree on are: (1) documents with higher comprehension difficulty; (2) longer documents; and (3) documents that are less cohesive. At least in so far as these three conditions are captured by the measures we have used, our results have been mixed. The correlation between entropy and disagreement confirms the third hypothesis, and provides a weakly useful practical predictor of disagreement. The relationship between document length and disagreement (our second hypothesis), if it exists, is too weak for our experiments to detect as significant. Most surprisingly of all, our first hypothesis, that difficult documents would provoke more disagreement, has not only failed to be confirmed, but in fact the reverse has been observed: it is easier documents that provoke the most disagreement.

As it seems intuitively hard to believe that it is in fact easily-comprehended documents that assessors disagree the most about, a more likely interpretation of our results is that the reading level measures are picking up some other aspect of document content, syntax, or representation that tends to provoke disagreement in assessors. An informal examination of disagreed-upon documents that attracted easy reading level scores, for instance, suggests that a disproportionate number of them are transcripts of spoken text—presidential debates, speeches, interviews, and the like. These tend to have short sentences, but diffuse topics, and may be difficult to read quickly. Further work is to determine whether there are other text metrics that can more directly and accurately target the aspects of a document that predict assessor disagreement.

Acknowledgments: This material is based in part upon work supported by the National Science Foundation under Grant No. 1065250. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. Measure-based metasearch. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, page 571, New York, New York, USA, 2005. ACM Press.

[2] P. Bailey, P. Thomas, N. Craswell, A. P. D. Vries, I. Soboroff, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of SIGIR*, SIGIR '08, pages 667–674. ACM, 2008.

[3] M. Bendersky, W. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 95–104. ACM, 2011.

[4] B. Carterette and I. Soboroff. The effect of assessor error on ir system evaluation. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 539–546. ACM, 2010.

[5] K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. *J. Am. Soc. Inf. Sci. Technol.*, 56(13):1448–1462, Nov. 2005.

[6] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 213–222, New York, NY, USA, 2012. ACM.

[7] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, Sept. 2000.

[8] J. Wang and D. Soergel. A user study of relevance judgments for e-discovery. *Proceedings of the American Society for Information Science and Technology*, 47:1–10, 2010.

[9] W. Webber. Re-examining the effectiveness of manual review. In *Proc. SIGIR Information Retrieval for E-Discovery Workshop*, pages 2:1–8, Beijing, China, July 2011.

[10] W. Webber, P. Chandar, and B. Carterette. Alternative assessor disagreement and retrieval depth. In *Proceeding 21st International Conference on Information and Knowledge Management - CIKM'12*, pages 125–134, 2012.

[11] W. Webber, B. Toth, and M. Desamito. Effect of written instructions on assessor agreement. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1053–1054, New York, NY, USA, 2012. ACM.