Nugget-Based Computation of Graded Relevance

Charles L. A. Clarke
Computer Science, University of Waterloo, Canada

ABSTRACT

We propose a simple method for assigning graded relevance values to documents judged during the course of a retrieval experiment. In making this proposal, we aim to avoid the potential for ambiguity and greater cognitive load associated with standard graded relevance judgments. Under our proposal, we first decompose a retrieval topic into a number of informational *nuggets*. For each document, a binary judgment is made with respect to each nugget. The ratio of relevant nuggets to total nuggets becomes the graded relevance value assigned to that document. To provide support for this idea, we turn to test collections created for the TREC Web Track. Along with the usual graded relevance judgments required by traditional effectiveness measures, these test collections include topic decompositions created for the purpose of evaluating novelty and diversity. By exploiting these test collections for our own purposes, we demonstrate a clear relationship between our proposed method and traditional graded relevance. In addition to supporting our proposal, our experiments suggest that informational nuggets can provide a unified approach to relevance assessment, supporting both traditional effectiveness measures and newer measures of novelty and diversity.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

search evaluation, graded relevance

1. INTRODUCTION

Accurate relevance judgments help retrieval experiments detect significant differences between systems, and provide better training data for learning ranking functions and tuning system parameters. Many classic retrieval experiments, including those conducted during the golden age of TREC, employed binary relevance judgments, where a document is either relevant or not [1, 10]. Under the relevance criteria established for these TREC experiments, a document is considered relevant if it contains any material related to the topic of a query, regardless of its quality or quantity.

Since the publication of the influential work of Järvelin and Kekäläinen in 2002 [6], retrieval experiments increasingly employ graded relevance judgements in an attempt to capture the relative quality of relevant documents. For example, documents might be judged with respect to five levels as "perfect", "excellent", "good", "fair", or "bad" [3]. Partly as a result of this work, graded relevance became standard for Web search evaluation [2, 3].

Carterette et al. [1] identify several concerns with graded relevance judgments. The cognitive load on assessors may increase with the number of levels, and with the complexity of their definitions. Moreover, the subjective nature of these definitions may make the choice of relevance level unclear, and changes to these definitions may impact effectiveness measures in unknown ways. Even with binary judgments the degree of assessor agreement can be under 70% [9, 10]. Graded levels may only exacerbate this problem.

Recognizing some of these concerns, Chapelle et al. [3] suggest a less subjective interpretation of graded relevance levels. They associate a probability $r(l)$ with each level l, where $r(l)$ is the probability that a user would consider a document at level l to be relevant. Unfortunately, they define an *ad hoc* formula for computing these probabilities, providing little justification for it.

Recent efforts to measure novelty and diversity [5,8] — including the TREC 2009-2011 Web Tracks [4] — take a different approach to relevance. For the TREC Web Track, query topics are decomposed into a number of *subtopics*, where each subtopic reflects a different aspect or interpretation of the overall topic. Assessors then make binary judgments with respect to each subtopic. For some of these topics, the associated query is ambiguous, and the subtopics reflect different interpretations of the query. For the remaining topics, the subtopics reflect different facets, or informational *nuggets*, associated with the topic. Essentially, each nugget provides a different perspective on relevance.

In related work, Pavlu et al. [7] propose nugget-based judgments as a method for partially automating the assessment process. While they adopt a different definition of a nugget, the core idea remains the same. To identify nuggets, they ask assessors "to find the smallest portion of text that constitutes relevant information in and of itself." These nuggets form the basis for a binary classifier, which determines relevance for unjudged documents.

EVIA'13, June 18, 2013, Tokyo, Japan

grade	number of relevant nuggets 0	1	2	3	4	mean
0	92.9%	6.5%	0.6%	0.0%	0.0%	0.08
1	0.0%	63.5%	28.1%	6.6%	1.8%	1.47
2	0.0%	25.0%	33.6%	27.4%	14.0%	2.32
3	0.0%	21.7%	18.3%	21.7%	38.3%	2.85

(a) TREC 2010 (22 faceted topics)

grade	number of relevant nuggets 0	1	2	3	4	mean
0	89.3%	8.3%	2.1%	0.3%	0.0%	0.13
1	0.0%	40.6%	37.5%	18.7%	3.1%	1.85
2	0.0%	25.9%	32.9%	30.0%	11.2%	2.27
3	0.0%	16.7%	42.8%	31.7%	8.8%	2.33

(b) TREC 2011 (41 faceted topics)

Figure 1: Graded relevance vs. number of relevant nuggets for TREC Web Track experiments.

2. GRADED RELEVANCE

It is intuitively reasonable that a more relevant document, when judged on a graded scale, would tend to be relevant to more nuggets. This idea holds regardless of whether we consider "more relevant" to mean "containing more relevant material" or "relevant to more of the user population".

Following the approach of Chapelle et al. [3] we treat graded relevance as a probability. Given a document d and a query q, we estimate its relevance value as

$$r = \frac{\text{number of nuggets for which } d \text{ is relevant}}{\text{total number of nuggets defined for } q} \quad (1)$$

Testing this simple idea requires only a test collection judged in both ways — in terms of graded relevance values and in terms of nuggets. Fortunately, the TREC 2010 and 2011 Web Tracks provide two such collections.

3. EXPERIMENT

For both the TREC 2010 and 2011 Web Tracks, the organizers created 50 new query topics. Of the 50 TREC 2010 topics, 22 are faceted topics, with subtopics suitable for use as nuggets. These faceted topics are decomposed into between two and six nuggets, with a mean of 4.2. The remaining topics are ambiguous, with subtopics reflecting different interpretations of the query, which we ignore.

Of the 50 TREC 2011 topics, 41 are faceted, reflecting a decreased emphasis on ambiguity that year. These faceted topics are decomposed into between two and five nuggets, with a mean of 3.4. For example, the query for TREC 2011 topic 122 ("culpeper national cemetery") is decomposed into nuggets related to the cemetery's location, history, and who can be buried there.

For both years, graded relevance is expressed on a four-point scale, with zero indicating that a document is not relevant. Although the assessment process differed slightly between the years, in both years the organizers enforced a requirement that a document receiving a non-zero graded relevance value must also be judged relevant to at least one nugget. However, they did not enforce the requirement that a document receiving a relevance grade of zero could not be judged relevant to any nugget.

Figure 1 compares graded relevance values against the number of relevant nuggets for all judged documents at TREC 2010 and TREC 2011. For each relevance level, the next five columns show the percentage of documents for different numbers of nuggets. The last column shows the mean number of nuggets for documents at that level. The relationship between relevance grade and the number of relevant nuggets supports our proposed method for computing graded relevance values.

4. CONCLUSIONS

We propose a simple approach to assigning graded relevance values. Our hope is that a series of binary judgements, rather than a single graded judgment, could reduce the cognitive load on an assessor and produce more consistent judgments. Our results, although preliminary in nature, provide support for this view, suggesting future work to further explore this idea. By building on the work of Pavlu et al. [7] it may be possible to partially automate the assessment process. The use of nuggets may provide a unified approach to relevance assessment, supporting both traditional effectiveness measures and newer measures of novelty and diversity.

5. REFERENCES

[1] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *30th ECIR*, pages 16–27, Glasgow, 2008.

[2] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *20th CIKM*, pages 611–620, Glasgow, 2011.

[3] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *18th CIKM*, pages 621–630, 2009.

[4] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. Voorhees. Overview of the TREC 2011 Web Track. In *20th TREC*, Gaithersburg, Maryland, 2011.

[5] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkann, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *31st SIGIR*, pages 659–666, Singapore, 2008.

[6] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.

[7] V. Pavlu, S. Rajput, P. B. Golbus, and J. A. Aslam. IR system evaluation using nugget-based test collections. In *5th WSDM*, pages 393–402, Seattle, 2012.

[8] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *34th SIGIR*, Beijing, 2011.

[9] M. D. Smucker and C. P. Jethani. Human performance and retrieval precision revisited. In *33rd SIGIR*, pages 595–602, Geneva, 2010.

[10] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *21st SIGIR*, pages 315–323, Melbourne, Australia, 1998.