Policy Capturing Models for Multi-faceted Relevance Judgments

Xiaoli Huang
College of Information Studies, University of Maryland, College Park, MD 20742. Email: xiaoli@wam.umd.edu.

Ryen W. White
Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742. Email: ryen@umd.edu.

We applied policy capturing and bootstrapping methods to investigate the relevance judgment process, with a particular focus on understanding how judges summarize an overall relevance judgment from five specific aspects of relevance. Our data come from relevance judgments made in the development of the MALACH (Multilingual Access to Large Spoken ArCHives) Speech Retrieval Test Collection. We developed a linear model for each of four relevance judges by regressing his or her overall judgments on the five specific relevance aspects. According to these models, different judges tended to assign different importance weights to different aspects. One of the linear models was applied to seven new judgment sets and was highly successful at predicting accurate overall judgments for the seven judgment sets.

Introduction

The purpose of this study is to better understand *relevance judgment* as a human judgment process. The understanding of topical relevance is generally oversimplified as merely "on topic", i.e., the document topic directly matches the user request. In contrast, every MALACH relevance judgment unit consists of five analytic topical relevance judgment scores – *direct*, *indirect*, *context*, *comparison* and *pointer* – and one holistic relevance judgment score, called *overall relevance* (see Huang and Soergel, 2004). These five types or aspects of topical relevance contribute to a richer concept of topical relevance; they play important roles in suiting different user situations and preferences. However, we do not fully understand the role played by different types of topical relevance and how each of them contributes to overall relevance. We analyzed quantitatively how relevance judges combine the individual aspects of relevance into one overall score, taking a first step towards addressing the issue.

We used *policy capturing*, a method that captures an individual's evaluative judgment process with algebraic models (Cooksey, 1996). Its purpose is to get insight of how people "weight, combine, or integrate information" (Zedeck, 1977) to reach a summary evaluative judgment. A policy capturing model can be used to predict an overall judgment from raw component judgments, a process called *Bootstrapping*. Bootstrapping works for a simple reason: people are good at picking out and weighing decision cues (in our case the five types of topical relevance) but bad at integrating them. Bootstrapping models systematically smooth random errors in the cue-to-judgment process (Dawes and Corrigan, 1974) and thus improve on human overall judgments.

Study

This study addresses the following research questions:

- What is the relative importance of the five specific rating aspects to the overall relevance judgment?
- How consistently do individual judges linearly relate overall judgments to the specific rating aspects?
- Does a linear model of the judge, compared to the judge themselves, provide greater judgmental validity?

In the summer of 2004, the MALACH project collected relevance judgments on a test collection of 400 testimonies (around 20,000 segments) for 50 topics from eight relevance judges (all graduate students in history or information studies). For each topic, two judges independently assessed segments, then met and discussed their judgments to decide on a final *adjudicated* judgment for each segment. Thus, for each topic-segment pair, there are two individual judgments and one final adjudicated judgment that serve as the validity evaluation baseline in this study.

This is a pilot study examining the relevance judgment processes of four of these judges on two topics:

A. "Doctors and Nurses in the Holocaust" yielding 1200 segments in the collection (Judge 1 and Judge 2)

B. "Life in the concentration camp" yielding 1048 segments in the collection (Judge 3 and Judge 4)

We developed a policy capturing model for each judge for the above two topics and used it to bootstrap overall relevance judgments. We looked at the validity of the model by evaluating the bootstrapped judgments against the baseline.

Findings and Discussion

We present the findings in four major analysis steps.

Step 1: Estimate the intercorrelations among the cues

In a standard policy capturing design, the cue values are experimentally manipulated in such a way that variable intercorrelations are minimized and hence the importance of independent cues can be accurately discerned. Although the five relevance rating dimensions are not completely orthogonal, the intercorrelations among them are low, ranging from +/– .005 (not significant) to – .369 (significant). The low intercorrelations allow further investigation and interpretation of the relative importance of each cue. *Pointer* dimension has the least correlations with other relevance dimensions, with only one significant correlation with *Direct* at a low level (r = .110). The strongest correlation (r = – .369) occurs between *Comparison* and *Direct* dimensions.

Step 2: Develop policy-capturing regression models

We developed a linear model for each of four judges by regressing his or her overall scores on the five relevance dimensions. With R^2 ranging from .770 to .896, all the multiple regression models derived from the data seem to effectively capture the underlying judgment patterns. *Direct* emerges as the most important dimension across the four judges and accounts for most variation in overall relevance. The judges differed in the importance they attached to the other dimensions (Table 1).

Table 1. Rank order of relevance aspects (best first)

Topic	Topic A		Topic B	
Judge	**Judge 1**	**Judge 2**	**Judge 3**	**Judge 4**
Order †	D,I,C,P,M	D,M,I,C,P	D,I,M,C,P	D,M,C,I,P

† D : Direct, I : Indirect, C : Context, M : Comparison, P : Pointer

Step 3: Evaluate the validity of judgment models

The validity of a policy model is defined as the correlation between the judgments predicted from the model and the baseline; the higher the correlation, the better the model. In our case, the base line is given by the adjudicated overall relevance scores. The regression model for each judge is used to generate a set of predicted overall scores, which we call bootstrapped scores. The validity coefficients for the bootstrapped scores (r^a) are then compared with validity coefficients for the raw overall scores assigned by the individual judges (r^b) (Table 2). The validity coefficients for the judges' raw overall judgments are high, ranging from .733 to .888. This indicates that the judges made consistent relevance judgments. The validity coefficients for predicted judgments are even higher, ranging from .856 to .922. In all four cases the bootstrapping models of the judges outperform the judges themselves.

Step 4: Apply the best model to test topics

To take this one step further, we used Judge 1's model (which has the highest r^a) to predict overall judgments for seven test topics (shown adjacent to Table 3), by bootstrapping from the adjudicated judgments for the five relevance dimensions. Judge 1 did not make judgments for these topics, which allows us to see whether the chosen model can predict overall judgments consistent with the baseline even when the baseline was defined completely by other judges. Table 3 shows that the judgmental accuracy of the Judge 1's model on these seven topics is high; with validity coefficients ranging from .883 to .961. We conclude that Judge 1's model is very effective in summarizing the cues to overall judgments. Moreover, the results also suggest that even though the adjudicated ratings are not ultimate accuracy baseline, they appear to be robust and reliable across different judges.

Table 2. Validity coefficients

Topic	Topic A		Topic B	
Judge	**Judge 1**	**Judge 2**	**Judge 3**	**Judge 4**
$r^a_{predicted}$	.922**	.856**	.868**	.888**
r^b_{raw}	.888**	.733**	.786**	.789**
Δ^c	+.034	+.123	+.082	+.099

$r^a_{predicted}$: correlation for judge models' predicted overall judgments
r^b_{raw}: correlations for individual judges' actual overall judgments
$\Delta^c = r^a_{predicted} - r^b_{raw}$; ** $p < 0.01$

Studies such as this allow us to understand and quantify the contribution of each relevance aspect and explore the interactions among them. This has both theoretical value in advancing the conceptualization of topical relevance and practical value in providing a direct step towards implementing an enriched relevance concept in IR systems.

Acknowledgments

This poster is based on a paper in the course *Theory of Decision and Choice*, instructor Tom Wallsten. The MALACH relevance judgment process was conducted under NSF grant IIS-0122466 and directed by Dagobert Soergel

References

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* New York: Academic Press.

Dawes, R. M. & Corigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95-106.

Huang, X. and Soergel, D. (2004). Relevance judges' understanding of topical relevance types: An explication of an enriched concept of topical relevance. *Proceedings of 67th ASIST Conference,* pp. 156-167.

Zedesk, S. (1977). An information processing model and approach to the study of motivation. *Organizational Behavior and Human Performance*, 18, 47-77.

MALACH website: www.clsp.jhu.edu/research/malach

Table 3. Validity coefficients

Topic	1	2	3	4	5	6	7
$r^a_{predicted}$	0.931**	0.924**	0.889**	0.911**	0.930**	0.883**	0.961**
N	663	441	578	565	366	351	466

$r^a_{predicted}$: correlations for assessor models' predicted overall judgments; ** $p < 0.01$; N: no. of judgments on topic

1. "Kindertransport"
2. "Nazis Eugenics Policies"
3. "Holocaust Survivors, Postwar, US"
4. "Post-liberation experience"
5. "IG Farden Labor Camps"
6. "Jewish Kapos"
7. "Art in Auschwitz"