Information Processing and Management 53 (2017) 1080–1102

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Relevance theory and distributions of judgments in document retrieval☆

Howard D. White

College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 18 December 2016
Revised 18 February 2017
Accepted 21 February 2017

Keywords:
Scale language
Cognitive effects
Processing effort

ABSTRACT

This article extends relevance theory (RT) from linguistic pragmatics into information retrieval. Using more than 50 retrieval experiments from the literature as examples, it applies RT to explain the frequency distributions of documents on relevance scales with three or more points. The scale points, which judges in experiments must consider in addition to queries and documents, are communications from researchers. In RT, the relevance of a communication varies directly with its cognitive effects and inversely with the effort of processing it. Researchers define and/or label the scale points to measure the cognitive effects of documents on judges. However, they apparently assume that all scale points as presented are equally easy for judges to process. Yet the notion that points cost *variable* effort explains fairly well the frequency distributions of judgments across them. By hypothesis, points that cost more effort are chosen by judges less frequently. Effort varies with the vagueness or strictness of scale-point labels and definitions. It is shown that vague scales tend to produce U- or V-shaped distributions, while strict scales tend to produce right-skewed distributions. These results reinforce the paper's more general argument that RT clarifies the concept of relevance in the dialogues of retrieval evaluation.

1. Introduction

When information retrieval systems are evaluated, judges assess retrieved documents for their relevance to queries, the simplest scale being "not relevant" or "relevant." A persistent finding in retrieval experiments is that scales with intermediate degrees of relevance produce distributions that are roughly U- or V-shaped. That is, whether the scales have few or many values, judges choose the low and high endpoints more frequently than the midpoints (Saracevic, 2007b: 2137), as illustrated at length by Janes (1993). However, in other articles, such as Sormunen (2002) and Lykke, Larsen, Lund, and Ingwersen (2010), the distributions of judgments on graded scales are right-skewed: from a high frequency at "not relevant," they simply decrease. This paper gives a cognitive account of factors underlying both kinds of distributions.

The argument is that the distributions depend not only on properties of the documents being assessed but also on judges' interpretations of the scale points, which are communications from researchers. The researchers may label and perhaps define the points, or they may simply let judges infer them as positions on a line. In either case, judges must expend cognitive effort to process the points as inputs, and the effort will vary, depending on properties of these inputs as communications.

My general claim is that judges tend to assign documents less frequently to scale points that cost them greater effort to process. That is, the more demanding the label, definition, or requirements of a point, the less judges choose it, resulting in

☆ *Disclaimer:* This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
E-mail address: whitehd@drexel.edu

http://dx.doi.org/10.1016/j.ipm.2017.02.010

highly non-uniform distributions of grades. White (2011) similarly equated greater effort with lower document frequencies to explain the skewing of citation distributions over time.

This line of thought is influenced by Dan Sperber and Deirdre Wilson's (1995) relevance theory, a major subfield of linguistic pragmatics brought most fully into information science (IS) by Harter (1992), Huang (2009), Saracevic (2007a), and White (2007a,b, 2009, 2010, 2011, 2014). Relevance theory (RT) and cognitive information science are mutually reinforcing. RT affords a clearer understanding of relevance in the dialogues of information retrieval (IR). Information science is a rich source of data for testing RT's explanatory powers.

The 50-plus experiments discussed here were found through topical searches, direct citation searches, and advice from experts. They carry forward Janes's (1993: 113) insight into what judges of documents do: "Determine, very quickly, if the document is really good or really bad. If so, say so (and the data appears to show that they don't much care exactly how really good or bad it is). If not, then more time and effort must be taken to determine how much of it is good, whether or not it is from a trustworthy source, addresses the right issues, is in the right language, is available and accessible, etc. The first of these processes is quick, relatively easy, and is done with confidence. The second is slower, less certain, and is done with more difficulty."

Studies mostly not covered by Janes are presented here as replicated tests of two hypotheses:

1. *U- or V-shaped distributions of relevance judgments tend to be associated with vague scales.* The vagueness is most problematic at the midpoints: the middle range is the muddle range (cf. Kulas & Stachowski, 2009). A sign of vagueness is that judges interpret scale points inconsistently; if they are asked *why* they placed a document at a certain point, as in Spink, Greisdorf, and Bateman (1998: 611) or Maglaughlin and Sonnenwald (2002: 335), their answers are all over the map. Vague scales also make for disagreements about document scoring, which have beset IR experiments from their earliest days (Saracevic, 2007b: 2134–2136).
2. *Right-skewed distributions of relevance judgments tend to be associated with relatively strict scales.* The definitions of points in these scales are somewhat less vague and become more exacting as they move rightward to the "most relevant" pole. Document counts at those points decline as the demands placed on judges increase. Strict scales are a reaction to the two-point scale defined by TREC, the Text REtrieval Conference (2006): "Only binary judgments ('relevant' or 'not relevant') are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)." Researchers such as Sormunen (2002) reject this formulation as too liberal. Accordingly, documents called "relevant" on TREC's binary scale in earlier studies have been reassigned to finer grades of relevance in, e.g., Sormunen (2002) and Järvelin (2013).

   Kekäläinen (2005: 283) writes: "Judging relevance liberally is fast. In graded assessment, extra work is required to specify the degree of relevance of each document." Researchers traditionally measure differences in effort by timing people's judgments: more difficult judgments take longer to make (Carterette & Soboroff, 2010; Smucker & Jethani, 2012; Wang, 2011). Retrieval evaluations in IS are analogous, though unintentional, tests of scale points. To differentiate the effort these points cost, I rely not on timing but on linguistic arguments buttressed by past empirical results. I developed effort-based accounts of the distributions on several prototypical scales and then sought uses of the same scales (or close variants) in other papers, none known in advance, to see whether distributions of the same shape occurred. The causal arrow runs from *language* to *frequencies*. Thus effort-based predictions can be confirmed or disconfirmed.

Judgments can of course be affected by cognitive factors other than the effort of processing scales. Research on such factors appears, for instance, in Ruthven, Baillie, and Elsweiler (2007) and Ruthven (2014), which also review older papers. The investigation reported here is not a meta-analysis of studies; it merely seeks, through RT, a parsimonious explanation of patterns in judgment frequencies at the level of entire experiments (or higher). Such aggregated patterns greatly simplify the causally complex distributions of judgments on individual queries. Distributions not fitting predicted shapes are noted.

2. Related work

2.1. New emphasis on effort

Several recent papers have analyzed user effort as a key variable in IR evaluation, especially as it bears on satisfaction with searches. Verma, Yilmaz, and Craswell (2016) measures and demonstrates the importance of three components of effort—reading a document, understanding it, and looking for specific information in it. According to a linked paper, Yilmaz, Verma, Craswell, Radlinski, and Bailey (2014), anyone must expend effort along these lines in the first stage of evaluating documents. But the authors also distinguish two kinds of judges. Further effort by "relevance judges" will consist simply in choosing the best grade for a document's degree of topical match with the query. By contrast, "real users" with acceptable first-stage results will put considerably more effort into a second stage, in which they extract the document's utility for their own purposes.

Jiang and Allan (2016) discusses user effort in processing a ranked list of documents to various depths. In past measures of retrieval quality, every ranked document has been treated as having the same cost to judge. The authors propose and test more realistic combined models that assign higher effort-costs to relevant documents in the list than to nonrelevant ones. Zuccon (2016) provides some measures that integrate document understandability (i.e., ease of reading) with topicality in