Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 44 (2008) 1117–1145
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Test theory for evaluating reliability of IR test collections

David Bodoff *

University of Haifa, Graduate School of Management, Jacobs Building, 6th Floor, 31905 Haifa, Israel

Received 15 July 2007; received in revised form 7 November 2007; accepted 12 November 2007
Available online 15 January 2008

Abstract

Classical test theory offers theoretically derived reliability measures such as Cronbach's alpha, which can be applied to measure the reliability of a set of Information Retrieval test results. The theory also supports item analysis, which identifies queries that are hampering the test's reliability, and which may be candidates for refinement or removal. A generalization of Classical Test Theory, called Generalizability Theory, provides an even richer set of tools. It allows us to estimate the reliability of a test as a function of the number of queries, assessors (relevance judges), and other aspects of the test's design. One novel aspect of Generalizability Theory is that it allows this estimation of reliability even before the test collection exists, based purely on the numbers of queries and assessors that it will contain. These calculations can help test designers in advance, by allowing them to compare the reliability of test designs with various numbers of queries and relevance assessors, and to spend their limited budgets on a design that maximizes reliability. Empirical analysis shows that in cases for which our data is representative, having more queries is more helpful for reliability than having more assessors. It also suggests that reliability may be improved with a per-document performance measure, as opposed to a document-set based performance measure, where appropriate. The theory also clarifies the implicit debate in IR literature regarding the nature of error in relevance judgments.

Keywords: Evaluation; Test collections; Test theory; Generalizabilty theory

1. Introduction

Information retrieval (IR) research has an impressive tradition of knowledge accumulation, in which high-performing methods are recognized and disseminated throughout the research community. A key ingredient in this impressive tradition is the availability of standard test collections and tasks. If an algorithm performs well on a test collection, it may imply that the model it implements will perform well in some more general sense. But in order to make that inference, the test must meet the twin challenges of internal reliability and external

* Tel.: +972 0 4 8249579; fax: +972 0 4 8249194.
E-mail address: dbodoff@gsb.haifa.ac.il

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.11.006

validity. External validity is the extent to which that performance would extend to other testing environments of interest, e.g. other tasks and data. A typical example of the question of a test collection's external validity is Soboroff's work "Do TREC Web Collections Look Like the Web?" (Soboroff, 2002). The work presented here focuses instead on the issue of internal reliability, which we will refer to simply as "reliability". Reliability is the extent to which a result reflects a real difference that is not due to chance. We note that the question of document "pooling" is another issue that has received much attention in the literature on IR evaluation. Although the connection to reliability and validity is not always explicit in that literature, the implicit goal of pooling is to achieve a good level of test reliability and validity while limiting the number-and expense-of manual relevance judgments.

The question of reliability can be addressed at the level of an individual performance comparison, or at the level of a whole test collection. For example, when an individual researcher finds that his/her model out-performs a baseline, he/she typically invokes a statistical test to show that the result is statistically significant and not due to chance (Hull, 1993). Our focus is not on the reliability of a given performance comparison, but the reliability of a whole test collection. We address the question, how do we define and measure the reliability of a test collection as a whole? The importance of the question is due to the fact that the accumulation of results in IR depends on the reliability of the test collections.

Our point of departure is the observation that the IR literature makes almost no use of Test Theory, which is a set of statistical tools specifically designed to rigorously define and measure the reliability of a test collection. Test Theory is in widespread use in the field of educational testing and in social sciences in general. It is used to measure the reliability of existing standardized tests, and to help plan the design of new ones. The basic insight of this paper is that the testing of algorithms on a set of test queries, is comparable to the testing of (human) students on a set of exam questions. Once this comparison is made, Test Theory suggests itself as a primary contender for the definition and calculation of the reliability of a test collection. There are many benefits from using Test Theory to assess test reliability. These benefits will become clear after we have reviewed the methods in current use and presented Test Theory as it applies to IR.

This paper is divided into two main parts. In the first part, which includes Sections 1–4, we introduce Test Theory, show how it applies to Information Retrieval test collections, and report specific results for a number of test collections and test designs. In the separable second part, which includes Sections 5,6, we explore two more advanced topics. Section 5 shows how to use Test Theory to consider the effect of documents, in addition to the effect of queries and assessors that we will analyze in Sections 1–4. Section 6 is more conceptual. It uses concepts from Generalizability Theory to illuminate a long-standing (previously implicit) problem in evaluation research, regarding how to model the effect of assessors on test reliability. This discussion clarifies the meaning of our own reliability results from Sections 1–4 as they relate to the effect of assessors, as well as clarifying the meaning previous research results.

The contributions of this paper are: (1) presentation of Test Theory and its proven set of tools that can be used to measure and improve IR test collection reliability. These tools complement the existing swap rate approach (reviewed in the next section) with a very different type of measure that also allows totally new types of analysis beyond the "mere" reporting of post-hoc reliability (2) presentation of results for a number of test collections. Results include both after-the-fact reliability and before-the-fact guidance for test collection designers (3) conceptual clarification of the variance that comes from assessors. Regardless of how one measures reliability – e.g. whether using a swap rate as in previous literature or Test Theory as proposed here – assessors are a main source of potential error in test collections, and we use ideas from Test Theory to illuminate a conceptual difficulty that has confronted previous work in this area.

1.1. Previous approach: the swap rate

We use the term "reliability" to refer to the extent to which a result is not due to chance. Our use of this term reflects a bias toward statistical methods, where this term has a technical meaning. However, in reviewing the literature, we will use "reliability" to refer also to non-statistical approaches such as the swap rate, which measure the effects of chance.

In a series of papers, Voorhees has analyzed reliability of TREC results using a data-driven concept of "swap rate". The swap rate of a set of test results can be calculated with respect to queries, relevance judges,