Predicting Relevance based on Assessor Disagreement

[Abstract] *

Thomas Demeester[1], Robin Aly[2],
Djoerd Hiemstra[2], Dong Nguyen[2], Chris Develder[1]
[1] Ghent University – iMinds, Belgium
[2] University of Twente, The Netherlands
tdmeeste@intec.ugent.be, r.aly@utwente.nl
{d.hiemstra, d.nguyen}@utwente.nl, cdvelder@intec.ugent.be

ABSTRACT

We present the Predicted Relevance Model (PRM): it allows moving from binary evaluation measures that reflect a single assessor's judgments, towards graded measures that represent the relevance towards random users.

1. INTRODUCTION

Evaluation of search engines relies on assessments of search results for selected test queries, from which we would ideally like to draw conclusions in terms of relevance of the results for general users. In practice, however, most evaluation scenarios only allow us to conclusively determine the relevance towards the particular assessor that provided the judgments. A factor that cannot be ignored when extending conclusions made from assessors towards users, is the possible disagreement on relevance, assuming that a single gold truth label does not exist.

We shortly describe the paper [1] that introduces the Predicted Relevance Model (PRM). The PRM allows predicting a particular result's relevance for a random user, based on an observed assessment and knowledge of the average disagreement between assessors. As a result, existing evaluation metrics designed to measure binary assessor relevance can be transformed into more robust and effectively graded measures that evaluate relevance towards a random user. The PRM also leads to a principled way of quantifying multiple graded or categorical relevance levels for use as gains in established graded relevance measures. Given a single set of test topics with graded relevance judgments, the PRM allows evaluating systems on different scenarios, such as their capability of retrieving top results, or how well they are able to filter out non-relevant ones. Its use in actual evaluation scenarios is illustrated on several information retrieval test collections.

2. THE PREDICTED RELEVANCE MODEL

The following definitions form the core of the PRM: (i) the user population of the search system under evaluation consists of individual users for whom a result is either relevant or non-relevant to a query, (ii) the assessors are part of the evaluation setup, and assign relevance labels to results according to well-described graded (or categorical) assessment levels i, and (iii) the disagreement parameters $p_{R|i}$ represent the probability that a random user would consider a particular result relevant (R), given the knowledge of an independent assessor judgment with level i.

The paper describes these concepts in detail, as well as provides a practical guide for calculating the disagreement parameters based on subsets of double judgments, and an extensive analysis of their properties.

Essential to the PRM are the distinction and the relation between the user model and the assessor model. For example, assessment levels on or above a threshold $i = \theta$ could define a scenario of binary user relevance. As an illustration, consider the task of counting the number N_R of relevant search results among a total set of N results. We denote the number of results assessed with relevance level i as n_i, such that $\sum_i n_i = N$. We can calculate N_R either by neglecting any disagreement between users and assessors (N_R^{bin}), or by taking the disagreement into account (N_R^{PRM}):

$$N_R^{\text{bin}} = \sum_{i \geq \theta} n_i, \quad N_R^{\text{PRM}} = \sum_i n_i \, p_{R|i}. \tag{1}$$

The expression for N_R^{bin} indicates the binary model based on the assessments alone, whereas N_R^{PRM} can be interpreted as the total *expected* number of relevant results for a random user [1]. The difference between both expressions is due to the assessor disagreement, and we show that it cannot be neglected in practice. A similar reasoning leads to the interpretation of metrics such as nDCG from the point of view of a random user, if the gain values are defined by the disagreement parameters, rather than chosen arbitrarily.

In a series of experiments based on existing evaluation collections, it is shown that the PRM leads to a robust evaluation of search engines with respect to several possible notions of binary user relevance.

3. REFERENCES

[1] T. Demeester, R. Aly, D. Hiemstra, D. Nguyen, and C. Develder. Predicting relevance based on assessor disagreement: analysis and practical applications for search evaluation. *Information Retrieval Journal*, 2015.

*A full version [1] of this paper has been accepted (Oct. 2015) for publication in Springer Information Retrieval Journal, special issue on *Information Retrieval Evaluation Using Test Collections*, with DOI: 10.1007/s10791-015-9275-x.