A new interpretation of average precision

Stephen Robertson
Microsoft Research,
7 JJ Thomson Avenue,
Cambridge CB3 0FB, UK
ser@microsoft.com

ABSTRACT

We consider the question of whether Average Precision, as a measure of retrieval effectiveness, can be regarded as deriving from a model of user searching behaviour. It turns out that indeed it can be so regarded, under a very simple stochastic model of user behaviour.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement, Experimentation

Keywords

IR evaluation, average precision, MAP.

1. INTRODUCTION

Average Precision (AP), or its mean over topics (MAP), is usually regarded as being a system-oriented measure, not based on a user model. The purpose of this paper is to provide a user model which does indeed give us AP as a measure. It is a very (grossly) simple user model, but it does accommodate some variant user behaviours in a probabilistic fashion. The fact that there does exist such a model is of interest, and further the model may suggest further developments of the measure.

The distinction between system- and user-oriented measures is misleading – all measures based on relevance are user-oriented to some degree. However, user models may be more or less explicit, and may or may not take account of differences across a population of users. If we can interpret a measure such as AP, which is normally regarded as system-oriented, in terms of an explicit user model (even better, a model of a population of users), this can only improve our understanding of what exactly the measure is measuring.

We make no assumptions here about whether 'relevance' itself is understood topically or subjectively. However, AP does assume that (a) relevance is binary, and (b) the relevance of a document to a request is independent of the other documents seen by the user. These are both of course over-simplifications in general.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

Recent work on measures motivated by models of user behaviour includes Rank Biased Precision (RBP) [5] and measures for structured document retrieval [6, 4, 3]. These will be discussed further below.

1.1 Notation

Precision at rank, P@n, is the number of relevant documents retrieved by rank n, divided by n. Reciprocal Rank, RR, is the reciprocal of the rank of the highest-ranked relevant document. For AP, we use the non-interpolated definition: the average of the P@n values at each *relevant* document in the ranking. This may be expressed as follows. We consider each rank position in turn and assess its contribution to AP. P@n depends on the documents ranked above n, but document d_n at rank n contributes to AP only if it is itself relevant. d_j has relevance $i_j \in \{0, 1\}$. For each $m < n$, define

$$\delta_{m,n} = \begin{cases} 1 & \text{if } i_m > 0 \textbf{ and } i_n > 0 \\ 0 & \text{otherwise} \end{cases}$$

Now we define AP as follows:

$$\text{AP}_n = \frac{1}{n} \sum_{m=1}^{n} \delta_{m,n} : \qquad \text{AP} = \frac{1}{R} \sum_{n=1}^{\infty} \text{AP}_n \tag{1}$$

AP_n (the contribution of d_n to AP) is zero if d_n is not relevant, so the latter sum can be over all documents in the collection. R is the total number of relevant documents. As usual we may assume that for incomplete rankings, the precision at an unranked relevant document is zero.

2. THE MODEL

Following Cooper in his proposal for the Expected Search Length measure [2], we envisage a user stepping down a ranked list until some stopping point. Stopping might for example be due to frustration or satisfaction, but Cooper concentrates on the notion of a satisfaction point. He assumes, and we follow this assumption, that satisfaction can only occur at a relevant document. In contrast to Cooper, we assume fully-ranked output with no ties, so the reason that Cooper introduced an expectation (which was to deal with ties in the ranking) no longer applies to us. However, we assume instead that we do *not* know the number of relevant documents that will satisfy the user. Rather we suggest a probabilistic model for this; this assumption will lead us back into taking an expectation over some set of probabilistically-defined events.

If we *know* where in a ranked list the user is going to stop, the actual rank order above and below this point is of no

consequence. With the above model of user behaviour, the user has seen everything above this point, but nothing below. Now a very obvious measure of effectiveness is the precision at this rank – like expected search length, this relates to the effort involved on the user's part in reaching the satisfaction point. We note that a view of precision in relation to user effort has also been taken in a number of recent proposals for measures in relation to structured document retrieval or web retrieval (see e.g. [6, 4, 3]). Here, probabilistic models of user behaviour are proposed, and measures are defined as expectations over these probabilistic event spaces. We follow a similar line; however, our model is somewhat simpler than those in not attempting to cover post-selection navigation.

2.1 Probabilistic user model for AP

Now we assume that with probability $p_s(n)$, the user's satisfaction point is the relevant document at rank n in the list (the $p_s(n)$s must sum to one; if d_n is not relevant, the probability is zero). This may be seen as a representation of a population of users with varying recall requirements, and consequently different stopping behaviours, described by these probabilities. We define a new measure, the expected precision observed by the user (the expectation over the above probabilistic event space) – Normalised Cumulative Precision, NCP:

$$\text{NCP} = \sum_{n=1}^{\infty} p_s(n)\text{AP}_n \tag{2}$$

using AP_n as given earlier. NCP lies between 0 and 1.

NCP requires the specification of the probabilities $p_s(n)$ for the set of ranked relevant documents. However, at this point we can define two simple versions. In the first, we assume a uniform distribution across all the relevant documents for the topic: we set all the non-zero $p_s(n)$ to $\frac{1}{R}$, where R is the number of relevant documents. This gives us exactly the usual non-interpolated AP:

$\text{NCP}_u = \text{AP}$ is the expected precision, given that our prediction of the user's stopping point is uniformly distributed over all the relevant documents for this topic.

This relates to (but is not quite the same as) other probabilistic interpretations of average precision, such as in [1].

This user model for AP is very simple and naïve. For one thing, it is probably much more likely that a user would stop after few relevant documents than after many. There will be a significant number of cases where users stop after just one relevant document, irrespective of how many there are in the collection. In fact we already have an exactly equivalent measure which addresses an appropriate version of the user model: namely, RR. In this version of the model, we simply set $p_s(1) = 1$ and all others to zero (here we do not need an expectation, since the user model is deterministic).

$\text{NCP}_1 = \text{RR}$ is precision, given that the user's stopping point is the highest-ranked relevant document for this topic.

2.2 Relation to Rank Biased Precision

The model on which RBP is based also involves a probability that the user will stop at a given rank. There are two main differences. First, for RBP this probability is assumed to be independent of the relevance of the document at that rank. Second, the structure of the stochastic process is different: the RBP probability is conditioned on the user reaching that rank. In our model, the satisfaction point is taken as a predetermined target: $p_s(n)$ is the unconditional probability that the user falls into the category of users who need to go as far as the nth relevant document.

We could advance arguments in favour of, or against, either stochastic user model. Both are very simplistic.

3. DISCUSSION

The fact that there exists a user model (albeit a simplistic one) for which AP is a natural effectiveness measure is of some interest – to our knowledge, no such model has been proposed before. In fact it is sometimes stated that there cannot be one – e.g. in [5] it is argued that 'there is no plausible search model that corresponds to MAP, because no user knows in advance the number of relevant answers in the collection they are addressing'. Actually, no such knowledge is required for the user model for AP described above. It is true that the probability of the user stopping at a particular relevant document depends on R; however, it is entirely plausible that such a correlation exists in the user population without such user knowledge (e.g. if topics that many people write about are on the whole more complex ones, that a user needs to read more about to understand).

The argument could provide the basis for a more elaborate model, by for example basing the set of $p_s(n)$ on some more sophisticated view of stopping behaviour. But even without that it does provide some insight into what AP is actually measuring and why it might be at the least a reasonable measure of IR effectiveness: one which assumes not a single prescriptive model of user behaviour, but rather allows for a mix of different behaviours in the population of users. In this respect it is just as well-justified from a user point of view as RBP, or indeed most other common measures.

4. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgements. In *SIGIR 2006*, pages 541–548. ACM Press, 2006.

[2] W. S. Cooper. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19:30–41, 1968.

[3] G. Kazai and M. Lalmas. Gain measures for the evaluation of content-oriented XML retrieval. *ACM Transactions on Information Systems*, 24:503–542, 2006.

[4] G. Kazai, B. Piwowarski, and S. Robertson. Effort-precision and gain-recall based on a probabilistic navigation model. In S. Dominich and F. Kiss, editors, *Studies in theory of information retrieval (Proceedings of ICTIR 2007)*, pages 23–36, Budapest, 2007. Foundation for Information Society.

[5] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *SIGIR 2007*, pages 375–382. ACM Press, 2007.

[6] B. Piwowarski and G. Dupret. Evaluation in XML information retrieval: expected precision-recall with user modelling (eprum). In *SIGIR 2006*, pages 260–267. ACM Press, 2006.