Evaluating Search and Hyperlinking: An Example of the Design, Test, Refine Cycle for Metric Development

David N. Racca, Gareth J. F. Jones
ADAPT Centre
School of Computing
Dublin City University, Dublin 9, Ireland
{dracca, gjones}@computing.dcu.ie

ABSTRACT

Designing meaninful metrics for evaluating MediaEval tasks that are able to capture multiple aspects of system effectiveness and user satisfaction is far from straighforward. A considerable part of the effort in organising such a task must often be devoted to selecting, designing or refining a suitable evaluation metric. We review evaluation metrics from the MediaEval Search and Hyperlinkiing task, illustrating the motivation behind metrics proposed for the task, and how reflection on results has led to iterative metric refinement in subsequent campaigns.

1. INTRODUCTION

It is a principle of MediaEval tasks that they should be built around a realistic use-case. This means that it is implicit in a MediaEval task that it should seek to evaluate participant submissions with respect to their effectiveness in performing the task, and that by implication that this should be related to a user's satisfaction with the actions of the system used in the participant's submission.

The objective of a MediaEval task will vary depending on the task itself. Measuring the success with which a particular system achieves its task objective can be complex, particularly in the case of temporal multimedia content [10]. For example, in conventional text information retrieval (IR) applications, items are often viewed as either relevant or non-relevant to the user's information need. While often much of such a document will not actually be relevant, it is generally deemed reasonable to label a document as either relevant or non-relevant without taking account of the the cost of identifying and extracting the relevant information from it. By contrast, in temporal media, the cost of identifying relevant content and extracting relevant information can be very significant. Thus, metrics typically make consideration of the specific points where relevant content begins and ends, and the cost, most often measured as the temporal distance, of locating this within a retrieved item. Further, temporal documents may be divided into segments in order to search for units with maximal proportions of relevant content to seek to promote their retrieval rank and improve content access efficiency. Measuring the multiple dimensions of relevance, retrieval rank and "cost" to access relevant content in a single metric presents many challenges.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2. EXAMPLE: SEARCH & HYPERLINKING

As a concrete example, let us consider the MediaEval Search & Hyperlinking (S&H) [3, 5, 4] task. We consider only the search sub-task which requires participants to find relevant video content from within a collection in response to a user query. The system is required to return a list of video segments (video ID, start time, end time), where start time suggests the beginning of a relevant portion of a video and end time suggests where this relevant content ends.

The task can be framed as an IR task and be evaluated by using the widely-adopted Cranfield paradigm for evaluating IR systems. In the context of the S&H task, this is implemented by first generating a pool of the top ranked retrieved segments (video ID, start time, end time) submitted by the participants for each query. Human assessors recruited through Amazon Mechanical Turk then judge the relevance of each individual segment in the pool with respect to its corresponding query. The set of segments judged relevant by the human annotators then forms the ground truth for the task, specifying for each query, which time spans in the video collection contain some relevant content.

2.1 User Models and Evaluation Metrics

The evaluation metrics used in the S&H search sub-task have all been based on standard Mean Average Precision (MAP). MAP models a user that scans a ranked results list from top to bottom looking for relevant items. MAP is calculated by computing the average of the precision at each rank where a relevant document is found for a query, and then computing the mean for a set of queries.

Standard MAP is not an appropriate measure for tasks like S&H where the cost of finding relevant information within a suggested relevant segment is non-negligible. Thus, various adaptations of MAP have been explored. Most of these take into account segment overlap or the distance to jump-in points, to compute the precision with which relevant content has been retrieved, and also reflect expected user effort to find and extract the relevant information.

Mean Generalized Average Precision (mGAP) [8, 10], is a variation of MAP which replaces simple binary relevance with a continuous function that penalises systems based on the distance from the ideal jump-in point to the beginning of a retrieved segment. In S&H 2014, three additional measures based on MAP were used: overlap MAP (MAP-over), binned MAP (MAP-bin), and tolerance to irrelevance MAP (MAP-tol) [1, 5]. While these metrics were designed carefully to measure performance in the S&H search task, subsequent analysis of results reveals weaknesses in all of them.

MAP-over rewards systems that return segments that overlap with some relevant content. As defined in [1], this measure presents various issues. First, a system receives extra credit if it returns multiple segments overlapping with the same relevant content. The metric therefore fails to acknowledge that most users will generally not want to see the same relevant content more than once. Furthermore, if a system retrieves more relevant items than the number of relevant segments in the ground truth, MAP-over can be ≥ 1 [7].

MAP-bin splits videos into bins of equal length. Bins overlapping with relevant segments are marked as relevant. A segment is considered relevant if its start time falls within a relevant bin. A system is therefore assumed to return a ranked list of bins and a user is assumed to watch the content of entire bins in the order given by their ranks. In contrast to MAP-over, in MAP-bin systems that return multiple jump-in points falling in the same relevant bin only get credit for its best-ranked instance. Analogously, systems that retrieve multiple jump-in points falling in the same non-relevant bin are penalised only once, even when checking every extra non-relevant bin may represent an additional effort for the user. Thus, a system that retrieves multiple jump-in points in the proximity of the intersection of two relevant bins is likely to obtain a higher MAP-bin score, because doing so would increase its chances of hitting more than just one relevant bin without receiving any extra penalty.

MAP-tol [2, 1] is a simplified form of mGAP which only rewards retrieved segments that start within a pre-defined tolerance window from unseen relevant content. In contrast to MAP-over and MAP-bin, MAP-tol successfully reflects the fact that users will not be satisfied if presented with content that they have seen before. However, MAP-tol equally rewards retrieved segments that point to large and short amounts of relevant content. It is thus more akin to standard MAP and not sufficiently informative of system behaviour.

Moving on from the variants of MAP introduced in 2014, for this year's search sub-task [4], we introduced a measure that estimates the user's effort in checking the relevance of each retrieved item and that does not reward duplicate results. User effort is measured in terms of the number of seconds that they must spend auditioning content, and user satisfaction in terms of the number of seconds of new relevant content that they can watch starting from a suggested jump-in point. This measure resembles Mean Average Segment Precision (MASP) [6], but differs from it in that precision is computed at fixed-recall points rather than at rank levels. Because of this similarity, we refer to it as MAiSP. We introduced two user models for MAiSP. MAiSP-ret assumes that the user watches the entire retrieved segment independently of whether the segment contains any relevant content. MAiSP-rel assumes that the user watches a retrieved segment until the end point suggested by the system in the case that no new relevant material continues thereafter, or until the last span of new relevant material is complete.

2.2 Correlation analysis

To compare the behaviour of these measures, we ran a series of retrieval experiments using the test collection used in the S&H 2014 search sub-task and computed the pairwise Pearson's r correlation between MAP-over, MAP-bin, MAP-tol, MAiSP-ret, and MAiSP-rel across 10,000 ranked lists produced with the Terrier IR platform [9]. Since most of the issues relating to the measures are more likely to be present in ranked lists that contain short and/or overlapping segments, we calculated correlation coefficients with the original set of 10,000 ranked lists and also with a modified version of the ranked lists that did not contain any overlapping segments in the results. Table 1 shows how the measures correlate when overlapping segments are removed from the ranked-lists. Most of the measures correlate relatively well in this case. However MAiSP-ret seems to be orthogonal to MAiSP-rel, MAP-bin and MAP-over, and to correlate negatively with MAP-tol. This is because MAiSP-ret is the only measure that assesses the quality of both the start and end time points of the retrieved segments. Table 2 shows correlations for the ranked lists that contain overlapping segments. MAP-over correlates negatively with most of the other measures, while MAP-bin correlates less strongly with MAP-tol and MAiSP than in Table 1, suggesting that these measures fail to penalise ranked lists containing duplicate results and that they therefore fail to reflect the users' preference against redundancy in the result lists.

	MAiSP rel	**MAiSP ret**	**MAP tol**	**MAP bin**	**MAP over**
MAiSP rel	1.0	-0.12	0.83	0.89	0.88
MAiSP ret	-0.12	1.0	-0.53	0.01	0.08
MAP tol	0.83	-0.53	1.0	0.78	0.74
MAP bin	0.89	0.01	0.78	1.0	0.86
MAP over	0.88	0.08	0.74	0.86	1.0

Table 1: Correlation between measures when overlapping segments are removed from the ranked-lists.

	MAiSP rel	**MAiSP ret**	**MAP tol**	**MAP bin**	**MAP over**
MAiSP-rel	1.0	-0.02	0.89	0.45	-0.47
MAiSP-ret	-0.02	1.0	-0.33	0.01	-0.33
MAP-tol	0.89	-0.33	1.0	0.51	-0.27
MAP-bin	0.45	0.01	0.51	1.0	0.39
MAP-over	-0.47	-0.38	-0.27	0.39	1.0

Table 2: Correlation between measures when the ranked-lists contain overlapping segments.

3. CONCLUSIONS

Designing evaluation measures for MediaEval tasks is often challenging. In tasks such as S&H, it is important to seek a measure of effectiveness which reflects the system's ability to find the necessary content and to maximise the satisfaction of the user in doing so. In the context of the S&H task, this essentially means minimising the user's effort in satisfying their information need. This note has shown how task evaluation measures can be refined over multiple editions of a task as the organisers come to better understand their task and reflect on its nature and its evaluation. From our experiences in the S&H task, it is important for task organisers to consider the necessary features of the evaluation metrics of the task, and to be open to reflecting on the strengths and weaknesses of the metric itself, as well as the calculated results when evaluating participant submissions.

4. ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Global Intelligent Content CNGL II project at DCU.

5. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, and G. J. F. Jones. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. Technical Report arXiv preprint arXiv:1312.1913, 2013.

[2] A. P. De Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO 2004 Conference Proceedings*, pages 463–473, Avignon, France, April 2004.

[3] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. F. Jones. The search and hyperlinking task at MediaEval 2013. In *Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[4] M. Eskevich, R. Aly, D. N. Racca, S. Chen, and G. J. F. Jones. SAVA at MediaEval 2015: Search and anchoring in video archives. In *Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, September 2015.

[5] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The search and hyperlinking task at MediaEval 2014. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[6] M. Eskevich, W. Magdy, and G. J. F. Jones. New metrics for meaningful evaluation of informally structured speech retrieval. In *Proceedings of ECIR 2012*, pages 170–181, Barcelona, Spain, 2012.

[7] P. Galušcáková and P. Pecina. CUNI at MediaEval 2014 search and hyperlinking task: Search task experiments. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, October 2014.

[8] B. Liu and D. W. Oard. One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 673–674. ACM, 2006.

[9] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in Terrier: a search engine for advanced retrieval on the web. *Novatica/UPGRADE Special Issue on Next Generation Web Search*, pages 49–56, 2007.

[10] P. Pecina, P. Hoffmannova, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF 2007 cross-language speech retrieval track. In *Proceedings CLEF'07*, pages 674–686, 2007.