SIGIR 2014 Workshop on Gathering Efficient Assessments of Relevance (GEAR)

Martin Halvey
Department of Computing and
Information Science,
University of Strathclyde
martin.halvey@strath.ac.uk

Robert Villa
Information School,
University of Sheffield
r.villa@sheffield.ac.uk

Paul D Clough
Information School,
University of Sheffield
p.d.clough@sheffield.ac.uk

Abstract

On July 11th 2014 the First Workshop on the Gathering Efficient Assessments of Relevance (GEAR 2014) was held as part of the SIGIR 2014 conference at the Gold Coast, Australia. An invited talk was given by Dr Nicola Ferro. Three full papers were presented, in addition to a design activity which lead to a lively discussion on gathering relevance assessments. This contribution discusses the events of the workshop.

1 Introduction

The Gathering Efficient of Relevance Workshop (GEAR) took place on 11th July 2014 as a full-day workshop in conjunction with ACM SIGIR'14. Evaluation is a fundamental part of Information Retrieval, and in the conventional Cranfield evaluation paradigm, sets of relevance assessments are a fundamental part of test collections, this served as a starting point for this workshop. This workshop revisited how relevance assessments can be efficiently created, seeking to provide a forum for discussion and exploration of the topic. Some of the themes of the workshop included:

- How the method of generating assessments, via conventional means or crowdsourcing, affects the judgments gathered, such as issues of assessor expertise, payment, etc.
- The process by which individuals, or groups of individuals, assess documents (text, image, video etc.) for relevance
- Issues relating to the effort required to generate relevance assessments for different types of topic, and different types of material (text, web, image, video, etc. and multiple languages)
- To revisit the concept of "relevance", from a practical, operational standpoint, for the purposes of IR evaluation.

Approximately 20 attendees registered for the workshop, the attendees were a mixture of people from academia, industry and government organisations. The schedule included one keynote presentation, followed by three paper presentations, a design activity which was discussed and a final open discussion. The workshop proceedings are available on CoRR. In the remainder of this report, we provide further details about the different parts of the workshop.

2 Keynote

The workshop began with a keynote presentation entitled “What do we mean by "Efficient Assessment”?” by Dr Nicola Ferro, University of Padua, Italy. Nicola began his keynote by introducing the typical life cycle of evaluation campaign, in particular focusing on the creation of pools by evaluation campaign organisers and assessment of relevance as provided by relevance. Nicola then moved on to discuss the meaning of efficiency, there are a number of definitions used in different fields, but in the context of information retrieval and evaluation campaigns efficiency can mean different things which in turn has different implications e.g. do we need faster assessments, do we need cheaper assessments, are we to demanding of assessors. As part of this Nicola highlighted the difference between “efficient assessments” and “efficient gathering of assessments”. As part of this distinction Nicola discussed a number of approaches for creating relevance assessments and the pros and cons for those approaches, the approaches discussed include interactive search and judge, shallow pools, preferences vs. absolute judgment, crowdsourcing, using no assessors, pseudo test collections etc. The keynote came back around to the assessors, and the question was posed as to whether we are asking too much of assessors and setting unrealistic expectations. The keynote concluded with some take home messages; that we should design with and for assessors, that there should be iteration between development of evaluation and that we might be overlooking vital information in how assessments are performed.

3 Paper Presentations

The succeeding session was dedicated to the presentation of research and position papers on the subject matter. Presentations were allocated thirty minutes each, with authors encouraged to only speak for 15 minutes to allow time for discussion. Three papers were chosen for presentation:

Aleksandr Chuklin, University of Amsterdam, presented “The Anatomy of Relevance: Topical, Snippet and Perceived Relevance in Search Result Evaluation” co-authored with Maarten de Rijke, University of Amsterdam [1]. In this position paper the authors highlight how two aspects of retrieval system performance can be affected by the presentation of results: result attractiveness i.e. perceived relevance and immediate usefulness of the snippets i.e. snippet relevance. The authors argue that perceived relevance may influence document discoverability and in turn render any ranking algorithm useless if attractive snippets lead to irrelevant documents or non-attractive snippets lead to relevant documents. As for snippet relevance, the authors believe that results pages with high snippet relevance may add to the total utility gained by the user even without clicking on those items. The authors demonstrate how collecting different aspects of relevance (topical, perceived and snippet relevance’s) can improve evaluation measures. Finally in this paper the authors discuss possible ways of gathering these relevance aspects using crowdsourcing and potential challenges arising from that.

Diego Molla, Macquarie University, presented “Document Distance for the Automated Expansion of Relevance Judgements for Information Retrieval Evaluation”, which was co-authored with Iman Amini, NICTA and RMIT, and David Martinez, University of Melbourne [3]. This paper outlined the use of a document distance-based approach to automatically expand the number of available relevance judgements when these relevance judgements are limited and reduced to only positive judgements. The paper compares the results of the author’s approaches on two different data sets: OHSUMED, based on medical research publications, and TREC-8, based on news feeds. The results of the evaluation indicate that evaluations based on these expanded relevance judgements are more reliable than those using only the judgements available initially, especially when the number of available judgements is very small.

Robert Villa, University of Sheffield, presented "Augmented Test Collections: A Step in the Right Direction" which was co-authored with Laura Hasler, University of Sheffield and Martin Halvey, University of Strathclyde [2]. In this position paper the authors argue that certain aspects of relevance assessment in the evaluation of IR systems are over simplified and that assessments as represented by qrels should be augmented to take account of contextual factors and the subjectivity of the task. The authors propose enhancing test collections used in evaluations with information related to human assessors and their interpretation of the task. This would provide a more realistic and user-focused evaluation, enabling us to better understand the evaluation process, the performance of systems and user interactions. The authors also outline some first steps that can be taken towards achieving this.

4 Design Activity

In the afternoon a design activity was presented to attendees. As part of the activity the attendees were presented with an example scenario around the National Fairground Archive (NFA). The NFA is a cultural heritage collection of photographic, printed, manuscript and audio-visual material covering all aspects of the culture of travelling show people, their organisation as a community, their social history and everyday life; and the artefacts and machinery of fairgrounds. The NFA collections has over 80,000 images in the photographic collection, in addition to audio and video material, journals and magazines, and nearly 4,000 monographs. The collection also includes a body of ephemera (programmes, handbills, posters, charters and proclamations, plans and drawings). The collection is accessed by a wide number of people from different backgrounds including historians, genealogists, engineers, performers etc. The original idea was to pose the attendees the task of determining the best way of collecting annotations or relevance assessments for this collection on a limited budget, with the approach and intended outcomes being left open ended in the hope of getting a breadth of approaches. Originally it was intended that participants would be split into groups, with each group consisting of attendees with differing interests, however, the workshop attendees stated a preference to have an open discussion instead. A number of issues and approaches were discussed, these included:

- Determining the information needs of different groups of end users, and also the frequency with which different groups look for particular types of information. As this some of the collection can be accessed online this could be achieved through query analysis, but as some of the collection is physical it would also be essential to interview librarians and users of the collection.
- Improving the current search interface was also discussed at length. This ties in with Nicola's keynote where he touched on developing interfaces to help users search but to also elicit more information from users.

5 Conclusion

This workshop was an attempt to revisit the notion of relevance and the approaches that are taken to gather relevance assessments for evaluations, but also beyond that in the more general sense. The audience for this workshop was gathered from a variety of backgrounds incorporating academia, industry and government. However, being a workshop organised in conjunction with the annual SIGIR conference, the majority of participants are from still the information retrieval community. Despite this a lot of the discussion did centre on the resources to capture, store and share relevance assessments. For evaluation campaigns, government and other small organisations this is a constraint on their abilities to share data. This was perhaps best encapsulated by a comment at the end of the workshop "big data is rich, small data is expensive"[1]. In the future, we intend to organise another

[1] https://twitter.com/leifos/status/487484922051850240

instalment of this workshop in conjunction with a different venue, perhaps in the area of digital libraries or human computer interaction to raise awareness but also to get different perspectives on gathering assessments efficiently but also effectively.

6 Acknowledgments

We would like to thank SIGIR for providing a venue for the workshop. This workshop was also supported financially by the Arts & Humanities Research Council UK, Digital Transformations in the Arts and Humanities (grant AH/L010364/1) and by the European Scientific Framework (ESF) under the Evaluating Information Access Systems framework (ELIAS).

7 References

[1] CHUKLIN, A. and DE RIJKE, M. The Anatomy of Relevance: Topical, Snippet and Perceived Relevance in Search Result Evaluation. *arXiv preprint arXiv:1501.06412*.

[2] HASLER, L., HALVEY, M., and VILLA, R. Augmented Test Collections: A Step in the Right Direction. *arXiv preprint arXiv:1501.06370*.

[3] MOLLÁ, D., AMINI, I., and MARTINEZ, D. Document Distance for the Automated Expansion of Relevance Judgements for Information Retrieval Evaluation. *arXiv preprint arXiv:1501.06380.*