The Anatomy of Relevance

Topical, Snippet and Perceived Relevance in Search Result Evaluation

Aleksandr Chuklin[*]   Maarten de Rijke

University of Amsterdam, Amsterdam, The Netherlands
a.chuklin, derijke@uva.nl

ABSTRACT

Currently, the quality of a search engine is often determined using so-called topical relevance, i.e., the match between the user intent (expressed as a query) and the *content* of the document. In this work we want to draw attention to two aspects of retrieval system performance affected by the *presentation* of results: result attractiveness ("perceived relevance") and immediate usefulness of the snippets ("snippet relevance"). Perceived relevance may influence discoverability of good topical documents and seemingly better rankings may in fact be less useful to the user if good-looking snippets lead to irrelevant documents or vice-versa. And result items on a search engine result page (SERP) with high snippet relevance may add towards the total utility gained by the user even without the need to click those items.

We start by motivating the need to collect different aspects of relevance (topical, perceived and snippet relevances) and how these aspects can improve evaluation measures. We then discuss possible ways to collect these relevance aspects using crowdsourcing and the challenges arising from that.

Categories and Subject Descriptors

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

1. INTRODUCTION

For decades the main evaluation paradigm for search engines was the Cranfield methodology [7]. In a typical setting of a TREC conference, the documents are evaluated by human raters who assign relevance labels based on their judgement about the relevance of the document to the user's topic of interest, expressed as a query. A graded relevance scale is typically used with topical relevance labels ranging from 0 to 4 or from *irrelevant* to *highly relevant*.

These relevance labels can be obtained either from trained experts or using a crowdsourcing approach. Either way, cases of disagreement have to be addressed, and those are usually treated as raters' mistakes, but may also arise from different interpretations of the user intent or the notion of relevance. In a traditional evaluation approach a single relevance label is chosen for each document-topic pair. These labels are then

[*]Now at Google Switzerland.

aggregated to SERP-level quality measures such as DCG [9] or ERR [3]. By using *additional* inputs from raters, we can (a) refine these quality measures and (b) better understand the performance of retrieval systems.

2. RELATED WORK

The idea to separate *perceived* and *topical* relevance was suggested by Chapelle and Zhang [2] while designing the DBN click model. Unlike earlier click models, it suggests that the likelihood of a user clicking a document depends not on the topical relevance of the document, but rather on its perceived relevance, since the user can only judge based on the result snippet. This idea was later picked up by Turpin et al. [12] who showed that while topical and perceived relevance are correlated, there is a noticeable discrepancy between them. They performed a simulated experiment by modeling the user click probability and showed that taking it into account would lead to substantially different ordering of the systems participating in a TREC Web Track.

The idea to separate out *snippet* relevance appears after the introduction of good abandonment [10]: cases when users abandon a search result page without clicking any results and yet they are satisfied. This may be due to the SERP being rich with instant answers [4], e.g., a weather widget or a dictionary box, or due to the fact that a query has a precise informational need, that can easily be answered in a result snippet [5]. In fact, as was shown by Stamou and Efthimiadis [11] a big portion of abandoned searches was due to a pre-determined behaviors: users came to a search engine with a prior intention to find an answer directly on a SERP. This is especially true when considering mobile search where the internet connection can be slow or the user interface is less convenient to use. We complement these works by arguing that good and relevant snippet does not necessarily lead to a complete good abandonment, but rather represents an aspect of utility gained by the user that is currently ignored.

3. APPLICATION TO EVALUATION

As was suggested by Carterette [1], many evaluation metrics, including DCG and ERR may be viewed as based on a click model. This was further refined by Chuklin et al. [6] where a recipe of converting any click model into a metric was presented:

$$uMetric = \sum_{k=1}^{N} P(C_k = 1) \cdot R_k, \qquad (1)$$

where R_k is the (topical) relevance of the k-th document in the ranking, and $P(C_k = 1)$ is the probability that the user will click on that document. Depending on the user model, the click probability may depend on attractiveness parameters. This is where we can use perceived relevance labels A_k (attractiveness). For example, for a metric based on the DCM model [8] we have:[1]

$$uDCM = \sum_{k=1}^{N} a(A_k) \prod_{i=1}^{k-1} (1 - a(A_i)s_i) \cdot R_k, \qquad (2)$$

where $a(A)$ is a list of parameters, one for each possible value of perceived relevance label A; s_i is another list of parameters, one for each value of the document position i.

Further, if we want to use snippet relevance labels S_k, we introduce a metric of the utility gained from the SERP itself similar to (1):

$$uMetric_S = \sum_{k=1}^{N} P(E_k = 1) \cdot S_k, \qquad (3)$$

where $P(E_k = 1)$ is the probability that the user examines the k-th document. Again, for DCM that would lead us to:

$$uDCM_S = \sum_{k=1}^{N} \prod_{i=1}^{k-1} (1 - a(A_i)s_i) \cdot S_k. \qquad (4)$$

To summarize, we showed that by collecting perceived, topical and snippet relevance we can refine system quality measures (eq. (1), (3)). To estimate the effect of this refinement one can compute correlations with online click metrics similar to [6] or with side-by-side comparison judgements collected using independent set of raters.

4. GATHERING JUDGEMENTS

Now that we have argued that perceived, topical and snippet relevance are potentially valuable dimensions of assessing system quality, how do we gather the required judgements? Firstly, we believe, that the *topical* relevance definition used by TREC raters is time-tested and hence can be used without modification. Secondly, *snippet* relevance can be treated as document topical relevance with document replaced by its snippet. We also need additional messaging for the raters explaining to them why the "documents" are so short to avoid undervalued scores. In order to prevent the raters from confusing this task with perceived relevance judgement, we may hide the fact that they are judging clickable snippets and just refer to them as short summaries.[2] Similar ratings were collected by Chuklin and Serdyukov [5], where three possible answers were offered to the raters: the snippet "answers the user question," "answers the question only partially," "does not answer the question." Third and finally, *perceived* relevance is a new task that has to be formulated by explaining to the rater the story of a web search and asking her if she would click this link in order to find the relevant information in the document. The snippet has to be shown without the context of the other snippets and without its placement on a SERP to avoid position and presentation biases.

These, then, are the challenges of gathering relevance judgements in the multi-aspect setting that we are proposing:

- How to make sure the raters do not confuse different tasks (topical, snippet, perceived relevance)?
- How do we treat special SERP items such as images, instant answers or interactive tools?
- What influence does the query category have on the difficulty of the task? For example, snippet relevance does not make sense for navigational queries.

5. CONCLUSION

This paper advocates for the need to review the notion of relevance in to order improve evaluation as well as understand the anatomy of relevance. We believe that after performing initial experiments and collecting feedback from the raters, we can address the challenges outlined above and derive a judgement procedure that will allow us to collect all three aspects of relevance, refine system performance evaluation and get deeper insights into the foundation of relevance.

REFERENCES

[1] B. Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *SIGIR*, 2011.

[2] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *WWW*. ACM, 2009.

[3] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*. ACM, 2009.

[4] L. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *WWW*, 2011.

[5] A. Chuklin and P. Serdyukov. Good abandonments in factoid queries. In *WWW*. ACM Press, 2012.

[6] A. Chuklin, P. Serdyukov, and M. de Rijke. Click model-based information retrieval metrics. In *SIGIR*, 2013.

[7] C. Cleverdon, J. Mills, and M. Keen. Aslib Cranfield research project – Factors determining the performance of indexing systems (Volumes 1 and 2). Techn. report, Cranfield University, 1966.

[8] F. Guo, C. Liu, and Y. Wang. Efficient multiple-click models in web search. In *WSDM*. ACM, 2009.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[10] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. In *SIGIR*, 2009.

[11] S. Stamou and E. N. Efthimiadis. Interpreting User Inactivity on Search Results. In *ECIR*, 2010.

[12] A. Turpin, F. Scholer, K. Jarvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *SIGIR*, 2009.

---

[1]A similar but more involved equation can be obtained for a metric based on the DBN model [2].

[2]It would be interesting to alter this clause in the rater's instruction and see how the outcomes change.