Comparing Click-through Data to Purchase Decisions for Retrieval Evaluation

Katja Hofmann, Bouke Huurnink, Marc Bron and Maarten de Rijke
ISLA, University of Amsterdam, The Netherlands
{k.hofmann, bhuurnink, m.m.bron, derijke}@uva.nl

ABSTRACT

Traditional retrieval evaluation uses explicit relevance judgments which are expensive to collect. Relevance assessments inferred from implicit feedback such as click-through data can be collected inexpensively, but may be less reliable. We compare assessments derived from click-through data to another source of implicit feedback that we assume to be highly indicative of relevance: purchase decisions. Evaluating retrieval runs based on a log of an audiovisual archive, we find agreement between system rankings and purchase decisions to be surprisingly high.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms: Experimentation, Human Factors

Keywords: Query log analysis, Evaluation

1. INTRODUCTION

Traditionally, information retrieval experiments use explicit relevance judgements. Annotators examine queries and candidate documents, explicitly judging which documents are relevant to a query. The use of explicit judgments is problematic: the judging process takes a lot of time, there can be wide interannotator variation [1], and explicit judging may not result in the same assessments that a user would make in a real search situation [6].

As an alternative to explicit relevance judgments, researchers have started to use click data from search engine transaction logs to infer relevance judgments for user searches [3, 5]. Click data can be collected unobtrusively, resulting in large numbers of judgments, and it reflects the search behavior of the original user. However, clicks are qualitatively different from explicit judgments, and agreement between the two is typically low. Kamps et al. [4] found large differences between system rankings based on explicit relevance assessments and those based on click data in web search.

We compare the use of click data for system evaluation to a related form of implicit relevance judgment — *purchase decisions*. One goal in commercial environments can be to rank items a user will buy as highly as possible. Here, purchase decisions would be a logical source of relevance judgments for evaluating system performance. Furthermore, if system evaluation based on clicks gives similar results to system evaluation based on purchases, then it would be possible to evaluate systems on a larger amount of data, as clicks are more plentiful than purchase decisions (in our data, the ratio between clicked items and purchased items is over 10 to 1), and commercially less sensitive.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

We address the following question: *do system rankings based on relevance judgments inferred from clicks agree with system rankings based on judgments inferred from purchase decisions?* We investigate this question by comparing the system rankings resulting from three sets of relevance assessments extracted from a large log of an audiovisual archive. The first two sets are based on queries that resulted in purchases — the first considers purchased items as relevant, the second clicked items. The third set is based on queries that did not result in purchases and considers clicked items as relevant. We simulate comparison of retrieval systems by generating 22 retrieval runs with different query weighting and term normalization schemes. Our goal here is to obtain a diverse set of runs, so that the sensitivity of evaluations using the different types of relevance assessments can be investigated. The runs are evaluated against the three sets of relevance assessments using standard retrieval measures and then ranked by performance.

We find that, in our setting, evaluation based on purchase decisions results in system rankings that are highly correlated with those based on click data. Rankings based on click data for queries that resulted in a purchase are close to identical to system rankings based on purchase decisions, while agreement with clicks on queries that did not result in a purchase is lower but still significant.

2. DATA AND METHODS

We first detail the collection and log data used in our study. Then we outline how queries and relevance assessments were derived, and how the retrieval runs were generated and evaluated.

We obtained click and purchase data from the Netherlands Institute for Sound and Vision, a large national audiovisual broadcast archive. To enable search, the archive indexes catalog entries describing the audiovisual documents in the collection. The archive primarily serves media professionals, who can search for and purchase audiovisual material for reuse in new productions. User interactions, including searches, result clicks, and purchases, are recorded in transaction logs; in this paper we use the data set described in detail in [2].

We use two sets of queries: (1) *purchase queries* that resulted in an order from the archive, and (2) *non-purchase queries* where results were clicked but nothing was ordered. Purchase queries are associated with both click data and purchase decisions. As an item must be clicked before it can be ordered, purchase decisions form a subset of the clicked results (in our data set overlap is 61%). Non-purchase queries are associated only with click data.

Result clicks and purchase decisions are mapped to queries in the archive transaction logs as follows: (1) identify the documents that were clicked for each query; (2) identify the documents that were purchased for each query; (3) collapse identical queries, and their associated clicked and purchased documents. Using this method

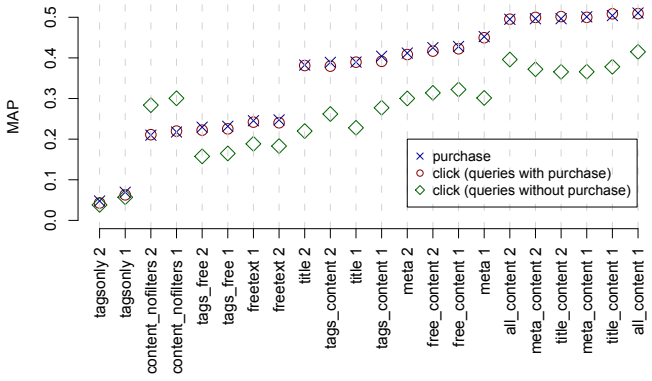

Figure 1: MAP per run and type of relevance assessment.

we extracted 13,506 unique purchase queries from the logs, which were associated with 28,761 clicks and 17,629 purchases. These queries were randomly split in half to obtain two non-overlapping sets of relevance assessments: one for evaluation using purchases, one using clicks on purchase queries. The third set of assessments was created using the 83,898 unique non-purchase queries contained in our data set. These were associated with 33,379 clicks.

Retrieval systems were created using the catalog entries maintained by the archive. We built two indexes using Lucene (http://lucene.apache.org/): for the first we preprocessed the collection using a standard tokenizer (index 1), for the second we also removed stop words and applied stemming (index 2). With the two indexes we generated 22 retrieval runs based on one or more of the following fields: *content* (all text associated with a document), *free(text)* (summary and description), *meta* (title and technical metadata), and *tags* (named entities, genre). By default we included the date filter and advanced querying options provided in the original search interface.

To assess retrieval performance we use mean average precision (MAP), mean reciprocal rank (MRR), and precision at 10 (P@10) using each set of relevance assessments. Systems are ranked according to each measure; agreement between rankings is measured using Kendall's τ rank correlation, and the number of pair-wise switches that would be required to turn one ranking into the other.

3. RESULTS AND DISCUSSION

We first summarize the retrieval scores obtained by our retrieval systems. Then we compare the system rankings obtained using the three sets of relevance judgments inferred from purchase and click data. Fig. 1 shows the performance in terms of MAP for all generated runs. System performance varies widely by run and assessment, ranging from 0.0382 (*tagonly 2*, clicks — no purchase) to 0.511 (*all_content 1*, clicks — with purchase). System rankings are similar for MRR (omitted due to limited space, absolute scores range from 0.049 to 0.538) and P@10 (0.008 to 0.097).

In terms of absolute scores, system performance is very similar when using purchase decisions and clicks from purchase data, even though these are obtained on different sets of queries. Differences are greater when looking at clicks from queries that did not result in a purchase. For the two runs *content_nofilter*, scores are substantially higher than when evaluating with queries that resulted in a purchase. For the remaining runs, scores are lower, with some systems changing ranks when evaluated on the different sets of queries. Despite differences in terms of absolute scores, a clear trend is visible: systems that score highly on one set tend to perform well on the other set too. Differences between evaluation scores based on purchases and clicks on non-purchase queries are systematic and indicate a qualitative difference between the two sets of queries. For example, queries that did not result in a purchase use date filters a lot less often (22% vs. 46%), which explains the performance jump on the *nofilter* runs. We think that queries that did not result in purchases are more exploratory in nature, while queries that did result in purchases include many known-item searches.

Table 1: Agreement between system rankings generated by *click* vs. *purchase* data according to standard evaluation measures. Agreement is calculated using Kendall's τ, and the number of pair-wise switches between ranked systems. All correlations are statistically significant with $p \ll 0.001$.

	purchases vs clicks from purchase queries		purchases vs clicks from non-purchase queries	
measure	τ	switches	τ	switches
MAP	0.974	6	0.766	54
MRR	0.948	12	0.766	54
P@10	0.991	2	0.775	52

For purchase queries (Table 1) system rankings using purchase decisions are highly correlated to those using clicks, with a rank correlation of 0.974 for MAP and similar values for MRR and P@10. In contrast, the correlation between system rankings using purchase decisions and those using clicks from non-purchase queries is lower at 0.77, but still statistically significant.

4. CONCLUSION

We investigated the use of purchase and click data for evaluating retrieval systems in a commercial setting. We found system rankings based on clicks to be close to identical to those based on purchase decisions when considering queries that resulted in a purchase. The high agreement between system rankings based on purchase decisions and those based on clicks is somewhat surprising as there is a marked difference in the size of the recall bases. While system ranking agreement is lower when evaluating systems with click data from non-purchase queries, it is still surprisingly high — especially in view of the low agreement that has been found between click data and explicit relevance assessments [4]. This may be due to the size of our data set and the professional nature of our users and their tasks; moreover, purchase decisions may be a better indicator for contextual relevance than explicit feedback.

Acknowledgements This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802.

5. REFERENCES

[1] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *JASIS&T*, 47(1):37–49, 1996.
[2] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. The search behavior of media professionals at an audiovisual archive: A transaction log analysis. *JASIS&T*, 2010. DOI: 10.1002/asi.21327.
[3] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05*, pages 154–161, 2005. ACM.
[4] J. Kamps, M. Koolen, and A. Trotman. Comparative analysis of clicks and judgments for IR evaluation. In *WSCD '09*, pages 80–87, 2009. ACM.
[5] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM '08*, pages 43–52, 2008. ACM.
[6] I. Ruthven. Integrating approaches to relevance. *New directions in cognitive information retrieval*, pages 61–80, 2005. Springer.