Enterprise Search: Learning to Rank with Click-Through Data as a Surrogate for Human Relevance Judgements

Colin Daly[1,2][a] and Lucy Hederman[1,2][b]
[1]*ADAPT Centre, Ireland*
[2]*School of Computer Science and Statistics, Trinity College Dublin, Ireland*

Keywords: Enterprise Search, Learning to Rank, Relevance Judgements, Click-Through Data.

Abstract: Learning to Rank (LTR) has traditionally made use of relevance judgements (i.e. human annotations) to create training data for ranking models. But, gathering feedback in the form of relevance judgements is expensive, time-consuming and may be subject to annotator bias. Much research has been carried out by commercial web search providers into harnessing click-through data and using it as a surrogate for relevance judgements. Its use in Enterprise Search (ES), however, has not been explored. If click-through data relevance feedback correlates with that of the human relevance judgements, we could dispense with small relevance judgement training data and rely entirely on abundant quantities of click-through data. We performed a correlation analysis and compared the ranking performance of a 'real world' ES service of a large organisation using both relevance judgements and click-through data. We introduce and publish the ENTRP-SRCH dataset specifically for ES. We calculated a correlation coefficient of ρ = 0.704 ($p<0.01$). Additionally, the nDCG@3 ranking performance using relevance judgements is just 1.6% higher than when click-through data is used. Subsequently, we discuss ES implementation trade-offs between relevance judgements and implicit feedback and highlight potential preferences and biases of both end-users and expert annotators.

1 INTRODUCTION

Enterprise Search is a federated store of workplace information with data gathered from multiple sources, such as intranets, document management systems, e-mail and social media (Kruschwitz and Hull, 2017; Craswell et al., 2005) and may also include the organisation's external-facing HTTP web servers (Hawking, 2004; Abrol et al., 2001).

Learning to Rank (LTR) is the application of supervised machine learning techniques for training a model to provide the best ranking order of documents for a given query (Li, 2011; Xu et al., 2020).

As with web search (WS), optimal ranking is also the major challenge for deployments of ES (Molnar, 2016; Craswell et al., 2005; Kruschwitz and Hull, 2017).

Krushwitz and Hull, in their 2017 book 'Searching the Enterprise' write that 'Search has become ubiquitous but that does not mean that search has been solved' (Kruschwitz and Hull, 2017). According to (Bentley, 2011), managers in the US, UK, Germany and France say that their internal enterprise search service falls short of expectations and that over half (52%) of surveyed users "cannot find the information they seek within an acceptable amount of time, using their own enterprise search applications".

ES differs from WS insofar as the content may be indexed from multiple databases (e.g. corporate directories) and intranet document repositories. ES may also feature alphanumeric searches for usernames, course codes, tracking numbers, purchasing codes or any datum specific to the organisation. In terms of ranking, this means that the determination of a 'good answer' for internet search is quite different than on the internet" (Molnar, 2016; Fagin et al., 2003). With ES, searches are often for known documents (such as this year's college calendar), or other well-structured objects (such as a person's contact details). This is sometimes referred to as a 'lookup search' (Marchionini, 2006) and is dependent on users 'recall and recognition' (Lykke et al., 2021). This is different to internet search, where a query is more likely to be 'exploratory' and complex (White and Roth, 2008).

Although most organizations are expected to de-

[a] https://orcid.org/0000-0001-7040-3305
[b] https://orcid.org/0000-0001-6073-4063

Daly, C. and Hederman, L.
Enterprise Search: Learning to Rank with Click-Through Data as a Surrogate for Human Relevance Judgements.
DOI: 10.5220/0012170200003598
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 1: KDIR*, pages 240-247
ISBN: 978-989-758-671-2; ISSN: 2184-3228

ploy an ES service, few have relevance judges (annotators) lined up to create training data. Consequently, the search service may operate somewhat 'relevance-blind' (Turnbull and Berryman, 2016). This scarcity motivates our investigation into the application of implicit feedback methods designed to replace annotations.

A click model is used to record end-user preferences on search results. In the field of WS, click models are considered as an effective approach to infer search relevance to a document for a given query (Wang et al., 2010).

An issue that the LTR community has not addressed is the extent to which click-through data, used as standard in WS, is also a good choice for ES. Specifically, we will test the hypothesis that the click-through rate is correlated with human relevance judgements and whether the correlation is statistically significant.

As the world around us changes, the relevance of a document for a given query also changes over time, leading to relevance drift (Moon et al., 2010). For example, the relevance of the term 'Ukraine' has changed dramatically over the past year. Enterprises cannot simply 'deploy and forget'. Periodic re-training of Learning to Rank models is required regardless of whether explicit or implicit feedback is used. An advantage of the implicit feedback approach is that it is not necessary to re-solicit judgements from expert annotators.

The contribution of this paper is to find out whether organizations, with an ES service, can safely dispense with expensive relevance judgements and rely instead on abundant log data. This is a question of particular significance for ES deployment projects with limited resources. A secondary contribution is the publication of a new LTR-formatted ES dataset that includes both human relevance judgements and corresponding click-through rates.

2 RELATED WORK

Since its introduction in 2005, Learning to Rank has become a hot research topic (Li, 2011) and has been used/promoted by commercial Web Search engine providers (Sculley, 2009; Qin and Liu, 2013). This growth was likely influenced firstly by Google making Learning to Rank more accessible with frameworks such as TensorFlow-Ranking (Wang and Bendersky, 2018), and secondly by Microsoft and Yahoo releasing datasets specifically designed to foster improvements for Learning to Rank algorithms (Liu et al., 2007; Chapelle, 2011).

An additional boost for Learning to Rank research was the publication of the Microsoft LETOR benchmark datasets that provide a basis for training and evaluating machine learning-based models (Qin et al., 2010). Much of current research on Learning to Rank pertains to comparing and evaluating the numerous ranking algorithms using the LETOR datasets, which are based on amorphous web content, fundamentally different from the multiple repositories, domain-specific nature of enterprise content (Mukherjee and Mao, 2004).

LTR involves supervised machine learning, and therefore a ground truth is needed to train the data. Much research has been carried out by commercial web search providers into harnessing click-through data and using it as a surrogate for relevance judgements (Kelly and Teevan, 2003; Joachims, 2002; Wang et al., 2010; Radlinski and Joachims, 2005; Jawaheer et al., 2010).

The general scarcity of academic studies on Enterprise Search environments stems from the difficulties of researchers gaining access to corporate environments (Cleverley and Burnett, 2019). A test collection based on Enterprise Search is hard to come by. An enterprise is not inclined to open its intranet to public distribution, even for research (Craswell et al., 2005). Furthermore, the corporation may decline permission to publish the results of any research carried out on Enterprise Search (Cleverley and Burnett, 2019; Craswell et al., 2005; Kruschwitz and Hull, 2017). Jawaheer has analysed the distinguishing characteristics between the various types of implicit and explicit feedback (Jawaheer et al., 2010). Table 1 outlines the specific differences between click-through log data and human relevance judgements, as applied to ES. The table shows how click-though feedback captures only positive user-preferences. Human judgements ought to have a greater accuracy as they also include negative feedback (e.g. via a Likert scale).

The literature pays scant attention to the amount of effort and expense involved in generating training data via the explicit feedback method. In the case of Enterprise Search, the annotators are likely to come from within the organization. This is because organizations are unlikely to release restricted intranet data to crowd-sourcing platforms such as Amazon Mechanical Turk (AMT). Moreover, the subject matter of the documents is specific to the enterprise and therefore only individuals with domain knowledge are well placed to volunteer relevance judgements.

A practical complicating factor for researchers of implicit feedback is that LTR-formatted datasets (see Figure 2 for an example) are generally published sep-

Table 1: Characteristics of human relevance judgements and click-through data as feedback for Enterprise Search.

Characteristic	Human	click-through
Accuracy	High	Lower
Abundance	Low	High
User preferences	+ve and -ve	+ve only
Domain knowledge	High	Low
Measurement relevance	Absolute	Relative
Principled Approach	Bureaucratic	Democratic

arately from associated click-through data. So while the 'Gov' corpus of Microsoft's LETOR 3.0 dataset identifies 64 features per document (Qin et al., 2010), no implicit feedback data is included.

According to a 2014 Google organic desktop search study, it was found that just 6% of end-users navigated to pages two and three. This contributes to a dramatic drop off in clicks for documents displayed on page 2 and inevitably leads to position bias (Petrescu, 2015). Position bias poses a well-known challenge to the integrity of implicit feedback and means that a direct correlation of click/non-click signals with positive/negative feedback respectively is confounded (Ai et al., 2018).

Similarly, explicit annotation may also be subject to bias. For example, 'organizational bias' occurs when factors such as strategic focus and team organization influence data selection to a point where selection is no longer based on individual merit (Dowsett, 2018; Zhang et al., 2019). This kind of bias can impact the relevance judgements of the organization's expert annotators.

3 METHOD

For the reasons outlined in §2, we could find no publicly available ES dataset that include both explicit and implicit feedback data that could be used for a correlation analysis and comparison. This research introduces a new learning to rank dataset, which has been extracted from the corpus of an enterprise website of a large third-level academic institution.

Firstly, we use the dataset to correlate explicit and implicit feedback methods. We then evaluate ranking performance using alternative 'ground truths' as input for the learning to rank method.

3.1 Data Collection

3.1.1 Enterprise Corpus

We have chosen the website of a large third-level education institution to build our corpus. Figure 1 shows a typical page from which we can identify and extract values from several example fields. The fields are signals of relevance, which are then coded to become ML features. For example, a combination of URL length, page hits and linkrank score can be an indicator that a document is a homepage (introductory page for the query term).

The corpus comprises about 67,000 documents (web pages, pdf documents, exam papers, invoice codes, people directory listings) crawled from a third-level educational institution's intranet and internet website. It includes fields such as URL, title, publication date and content (body), as shown in Figure 1. Apache Solr (Białecki et al., 2012) is the technology used to host and index the corpus, which was populated using the Apache Nutch crawler (pages on the site were crawled June 12th 2022). The site includes an ES service that receives about 7,000 queries daily. In addition to features extracted from the page itself, further features (such as hitcount and query-dependent click-through rate) are extracted from 180 days worth of the Apache web server's log files.

3.1.2 Dataset

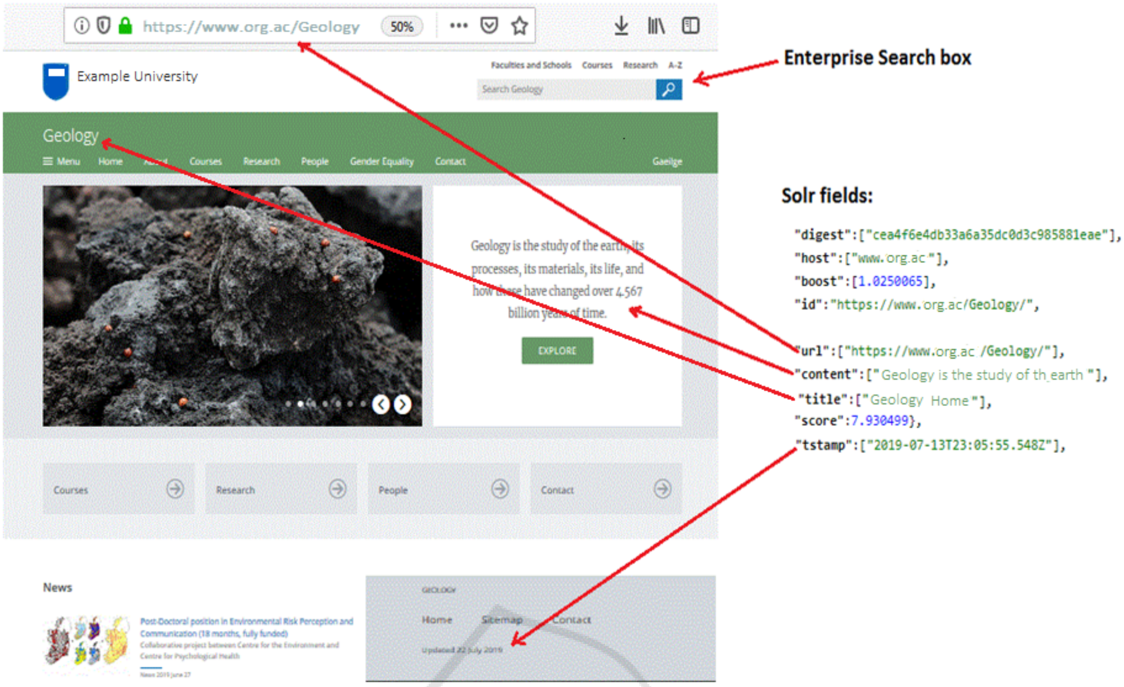
From this corpus, we develop a small-scale dataset that consists of 20 queries and 2544 Query-Document (Q-D) pairs, with manually annotated relevance judgements. The queries were selected to be representative of typical search requests (or clusters of requests) as extracted from Apache Solr log data. The dataset is presented in the LETOR format, which includes relevance judgements in the first column and an associated feature vector array. We name our anonymized dataset 'ENTRP-SRCH'. Table 2 compares its properties against popular LTR datasets. ENTRP-SRCH is publicly available for download at https://github.com/colindaly75/ES-LTR-Implicit-Explicit-Correlation.

Table 2: Comparison of the properties of popular LTR datasets and our ENTRP-SRCH dataset.

	Microsoft	Yahoo!	ENTRP-SRCH
Pub. Year	2010	2010	2022
Docs	3771K	883K	2544
Queries	31531	36251	20
Doc/Query	119	24	127
Features	136	700	8
No. of click-through values recorded	None avail.	None	375
Corpus Type	WS	WS	ES

3.1.3 Human Relevance Judgements

Web authors and domain experts were asked to judge the relevance of a document for a given query. Fifteen university staff members, who each maintain a sub-section of the university website were engaged



```
"digest":["cea4f6e4db33a6a35dc0d3c985881eae"],
"host":["www.org.ac"],
"boost":[1.0250065],
"id":"https://www.org.ac/Geology/",

"url":["https://www.org.ac /Geology/"],
"content":["Geology is the study of th earth "],
"title":["Geology  Home"],
"score":7.930499},
"tstamp":["2019-07-13T23:05:55.548Z"],
```

Figure 1: The relationship between Enterprise Search on a typical university website and retrieved fields used to construct features for a learning to rank dataset.

for this purpose. The annotators have great familiarity with the content as they are the individuals tasked with publishing on behalf of their department or faculty.

Annotators were asked to limit the number of '5's awarded. A smaller number of highly relevant documents are offered for users' attention, thereby mitigating the effects of position bias (Wang et al., 2018). Since the number of 'highly relevant' documents is smaller than the number of 'moderately relevant' documents, the judgements scores do not follow a normal distribution (Table 3).

Table 3: Distribution of human relevance judgements in the ENTRP-SRCH dataset.

Relevance Label	Interpretation	Number	Percentage
5	highly relevant	147	5.78%
4	relevant	184	7.23%
3	moderately relevant	359	14.17%
2	irrelevant	1639	64.45%
1	utterly irrelevant	214	8.42%

3.1.4 Click-Through Data

The click-through rate (CTR) is defined as the percentage of the number of clicks to the number of impressions (Chapelle, 2011). When our end-user submits a query to our search engine, he/she is presented with a list of documents. Our click model simply records the ordered list and which result is clicked.

A high CTR is a good indication that users find the document within the search results as helpful and relevant for the given query.

3.1.5 Learning to Rank Features

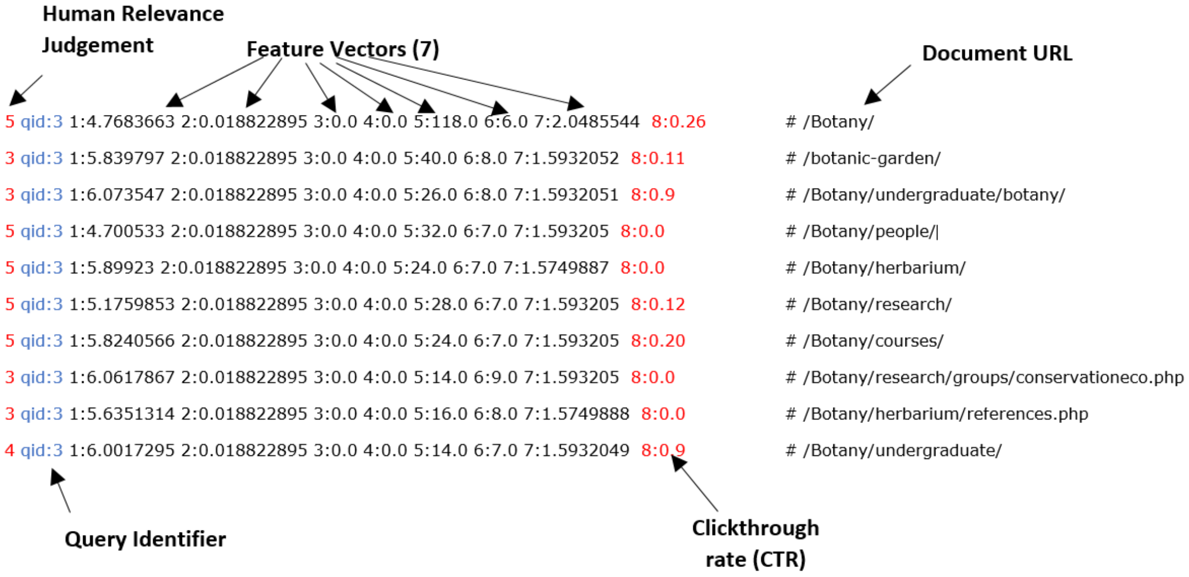
A Learning to Rank can be scaled to include any number of features. Features implemented in our ENTRP-SRCH dataset include BM25 (Robertson et al., 1995), documentRecency (last modification date), rawHits (a measure of document popularity), urlLength (number of terms in url path hierarchy), linkRank (Kim et al., 2010) (based on a web graph, this link analysis algorithm is similar to Google's PageRank) and clickThru (CTR score).

The relevance judgements per query are then combined with the feature vector matrix to create our dataset. Figure 2 shows our dataset features presented in the LETOR format (Qin et al., 2010).

3.2 Experiments

3.2.1 Correlation

We plot the correlation between the explicit human relevance judgements and the calculated CTR score. We examine the causes of any disparity between the annotator's judgements and the click-through data as recorded by end-user search preferences.



```
5 qid:3 1:4.7683663 2:0.018822895 3:0.0 4:0.0 5:118.0 6:6.0 7:2.0485544  8:0.26      # /Botany/
3 qid:3 1:5.839797 2:0.018822895 3:0.0 4:0.0 5:40.0 6:8.0 7:1.5932052  8:0.11         # /botanic-garden/
3 qid:3 1:6.073547 2:0.018822895 3:0.0 4:0.0 5:26.0 6:8.0 7:1.5932051  8:0.9          # /Botany/undergraduate/botany/
5 qid:3 1:4.700533 2:0.018822895 3:0.0 4:0.0 5:32.0 6:7.0 7:1.593205  8:0.0           # /Botany/people/|
5 qid:3 1:5.89923 2:0.018822895 3:0.0 4:0.0 5:24.0 6:7.0 7:1.5749887  8:0.0           # /Botany/herbarium/
5 qid:3 1:5.1759853 2:0.018822895 3:0.0 4:0.0 5:28.0 6:7.0 7:1.593205  8:0.12         # /Botany/research/
5 qid:3 1:5.8240566 2:0.018822895 3:0.0 4:0.0 5:24.0 6:7.0 7:1.593205  8:0.20         # /Botany/courses/
3 qid:3 1:6.0617867 2:0.018822895 3:0.0 4:0.0 5:14.0 6:9.0 7:1.593205  8:0.0          # /Botany/research/groups/conservationeco.php
3 qid:3 1:5.6351314 2:0.018822895 3:0.0 4:0.0 5:16.0 6:8.0 7:1.5749888  8:0.0         # /Botany/herbarium/references.php
4 qid:3 1:6.0017295 2:0.018822895 3:0.0 4:0.0 5:14.0 6:7.0 7:1.5932049  8:0.9         # /Botany/undergraduate/
```

Figure 2: The learning to rank dataset. Each row represents a query-document pair. This dataset contains both explicit relevance judgements (first column) as well as the calculated click-through rate (in red). In the example, the query with id '3' includes a value of '8:0.26' for the first document, meaning that feature 8 has CTR of 26%.

3.2.2 Ranking Performance

A ranking model is generated using the XGBoost implementation of the LambdaMART list-wise ranking algorithm (code on GitHub). Since version 6.5, Lucene Solr has a built-in *contrib* module called *Learning to Rank (LTR)* which can be used to re-rank the top-N retrieved documents using trained machine learning models. Hence our experience of integrating the trained model was relatively straightforward.

To compare ranking performance, we applied two different 'ground truths' to the training, test and validation datasets and calculated the nDCG score for each: -

- Explicit feedback (i.e. human relevance judgements) is the first ground truth to be tested and is the one traditionally associated with learning to rank.
- Secondly, we use implicit feedback (CTR score) as an alternative (surrogate) ground truth.

In the field of ranking performance, the Normalized Discounted Cumulative Gain(nDCG) metric is typically used (Tax et al., 2015). nDCG is often described as a rank-aware metric (it credits the fact that some documents are quantitatively 'more' relevant than others).

4 EVALUATION

4.1 Correlation

The first step of the evaluation is to establish to what degree of accuracy do clicks correspond to explicit judgments of a document for a given query. Figure 3 is a strip plot that suggests a correlation between the human relevance judgements and the CTR. We see that those documents, that have been labelled with a higher relevance score by the expert human annotators for a given query, also tend to be more clicked by end users.

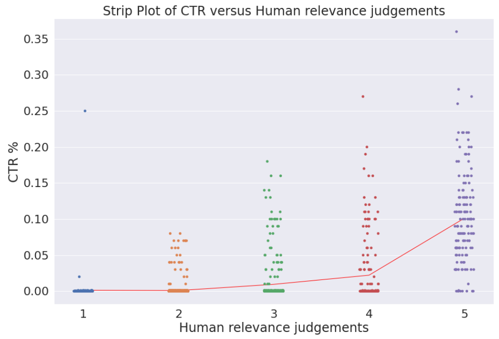


Figure 3: A strip scatter plot showing points of correlation between click-through rate (CTR) on the y-axis and human relevance judgements on the x-axis. Those documents that received a higher relevance judgement tend to have recorded more click-through activity.

As per the hypothesis outlined in §1, we calculate the correlation coefficient as a measure of the strength of the relationship between CTR and ordinal human relevance judgements. Since the human relevance variable is ordinal in nature (a Likert scale) with a non-normal distribution of judgements, Spearman's rho is the appropriate correlation metric. Using the same data that was used to plot figure 3, we calculate a correlation score of $\rho = 0.704$. Most statisticians consider a score of 0.7 or above as a 'strong correlation' (Akoglu, 2018). A significance test using the paired t-test gives a p-value of less than 0.01, proving that the Spearman correlation is statistically significant.

4.2 Correlation Analysis

Figure 3 shows a slight divergence from a linear correlation plot (i.e. a straight line), insofar as there seems to be an excess of points 'under the line'. For example, where the human relevance judgement score is 4 ('relevant', but not 'highly relevant'), there are many points with a low CTR rate. Analysis of the respective documents suggests that this non-linearity may be caused by the fact that few end users ever navigate to pages 2 or 3 of the search results (the ranking model displays a preponderance of 'highly relevant' documents on Page 1). This is in line with the 'position bias' confounding problem outlined in §2 (i.e. reports that less than 6% of end users navigate beyond Page 1 of the Search Engine Results Page).

End users are not the only feedback party subject to bias. 'Organizational bias' occurs when factors such as strategic focus and team organization influence data selection to a point where selection is no longer based on individual merit. We see some instances of disagreement over pages judged to be 'irrelevant' by expert annotators but nonetheless have been awarded a high click-through rate by end users. For example, the homepage of a VIP / celebrity individual in the organization has been annotated as 'irrelevant' (for the given query). This judgement dampens favouritism and diminishes the democratic preferences of end users. In this case, the experts are subject to organizational bias, where they overlook an individual's popularity in favour of a bureaucratic approach.

A further example of a pronounced contradiction between annotators and end-users was detected. One of the query terms in our training dataset is 'english'. The annotator for the english query is employed by the department of english and therefore assigned preferential judgements based on his department's perspective. The end-users, many of whom are prospective students from non-English speaking countries, were less interested in literature and instead wanted to ascertain minimum English language requirements for entry to the student register.

4.3 Comparing Fit of Features

If we alternate the ground truth in our LTR model, such that either CTR scores or human relevance scores are used to train the data, there will be a resultant change to how well the features combine to 'fit'.

The nDCG values for both feedback methods are shown in Table 4 and graphically represented in Figure 4. This shows that LTR model's custom features, as listed in section 3, are better at matching explicit rather than implicit feedback. This is to be expected, as the features were initially engineered to match the requirements of human relevance judgements.

Table 4: Comparison of ranking performance (nDCG) for relevance judgements (explicit annotator feedback) versus query preferences extracted from click-through (implicit feedback).

Cutoff	Relevance Judgements	Click-through
ndcg@1	**1**	0.997
ndcg@3	**0.987**	0.970
ndcg@5	0.941	**0.964**
ndcg@10	0.787	**0.963**
ndcg@20	0.567	**0.853**
ndcg@100	**0.335**	0.271
ndcg@200	**0.313**	0.061

5 CONCLUSIONS AND FUTURE WORK

This paper evaluated using two approaches to feedback as training data for generating ranking models in the domain of Enterprise Search.

A new learning to rank dataset, ENTRP-SRCH, was generated from the intranet and internet-facing parts of a large third-level institution.

We plotted the correlation between the human relevance judgments and respective click-through rates for a given query. Outliers and irregularities on the correlation plots are explained by end-user 'position bias' and annotator 'organizational bias'. We hypothesized and proved that there was a strong correlation between implicit and explicit feedback.

Furthermore, by alternating implicit and explicit feedback as ground-truth in our LTR model, we achieved similar nDCG scores for our ranking model based on our custom features.

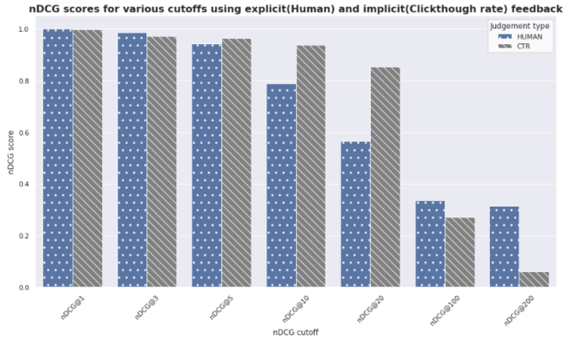


Figure 4: A bar chart showing the performance differences between explicit and implicit feedback for various nDCG cutoffs. For nDCG@3, the human relevance judgements are just 1.6% higher than those recorded using click-through feedback. For nDCG@5, nDCG@10 and nDCG@20, the use of CTR as ground truth achieves higher scores than human judgements.

Future work may include mitigation of the identified bias in both approaches, e.g. by applying an inverse propensity score or introducing more diversity to annotator selection.

Enterprise content is diverse and different for every organisation. The generalisability of the ENTRP-SRCH dataset is therefore limited. However, since click-through feedback is cheap and abundant compared to human relevance judgements, our (correlation and ranking performance) findings for our organisation may present a crucial cost-saving opportunity to other organisations considering which type of feedback approach they should adopt for learning to rank in the context of Enterprise Search.

ACKNOWLEDGEMENTS

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 P2 at the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

REFERENCES

Abrol, M., Latarche, N., Mahadevan, U., Mao, J., Mukherjee, R., Raghavan, P., Tourn, M., Wang, J., and Zhang, G. (2001). Navigating Large-Scale Semi-Structured Data in Business Portals. In *International Conference on Very Large Data Bases*, VLDB '01, page 663–666, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ai, Q., Mao, J., Liu, Y., and Croft, W. B. (2018). Unbiased Learning to Rank: Theory and Practice. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '18, page 1–2, New York, NY, USA. Association for Computing Machinery.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91.

Bentley, J. (2011). Mind the Enterprise Search Gap: Smartlogic Sponsor MindMetre Research Report.

Białecki, A., Muir, R., Ingersoll, G., and Imagination, L. (2012). Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*, page 17.

Chapelle, O. (2011). Yahoo! Learning to Rank Challenge Overview.

Cleverley, P. H. and Burnett, S. (2019). Enterprise search and discovery capability: The factors and generative mechanisms for user satisfaction:. *Journal of Information Science*, 45(1):29–52.

Craswell, N., Cambridge, M., and Soboroff, I. (2005). Overview of the TREC-2005 Enterprise Track. In *TREC 2005 conference notebook*, pages 199–205.

Dowsett, C. (2018). It's Time to Talk About Organizational Bias in Data Use.

Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A., and Williamson, D. P. (2003). Searching the workplace web. *Proceedings of the 12th International Conference on World Wide Web, WWW 2003*, pages 366–375.

Hawking, D. (2004). Challenges in Enterprise Search. In *Proceedings of the 15th Australasian Database Conference - Volume 27*, ADC '04, page 15–24, AUS. Australian Computer Society, Inc.

Jawaheer, G., Szomszor, M., and Kostkova, P. (2010). Comparison of implicit and explicit feedback from an online music recommendation service. *Information Heterogeneity and Fusion in Recommender Systems, HetRec 2010*, pages 47–51.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference. *ACM SIGIR Forum*, 37(2):18–28.

Kim, Y., Son, S. W., and Jeong, H. (2010). LinkRank: Finding communities in directed networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 81(1).

Kruschwitz, U. and Hull, C. (2017). Searching the Enterprise. *Foundations and Trends® in Information Retrieval*, 11(1):1–142.

Li, H. (2011). A Short Introduction to Learning to Rank. *IEICE Transactions*, 94-D:1854–1862.

Liu, T.-Y., Xu, J., Qin, T., Xiong, W., and Li, H. (2007). LETOR: Benchmark Datasets for Learning to Rank. *SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, 1(Lr4ir):3–10.

Lykke, M., Bygholm, A., Søndergaard, L. B., and Byström, K. (2021). The role of historical and contextual knowledge in enterprise search. *Journal of Documentation*, 78(5):1053–1074.

Marchionini, G. (2006). Exploratory search. *Communications of the ACM*, 49(4):41–46.

Molnar, A. (2016). Google Search Appliance Retirement Explained. Technical report, Search Explained, Diosd.

Moon, T., Li, L., Chu, W., Liao, C., Zheng, Z., and Chang, Y. (2010). Online learning for recency search ranking using real-time user feedback. *International Conference on Information and Knowledge Management, Proceedings*, pages 1501–1504.

Mukherjee, R. and Mao, J. (2004). Enterprise Search: Tough Stuff. *Queue*, 2(2):36.

Petrescu, P. (2015). Google Organic Click-Through Rates in 2014.

Qin, T. and Liu, T.-Y. (2013). Introducing LETOR 4.0 Datasets. *Microsoft Research Asia*.

Qin, T., Liu, T. Y., Xu, J., and Li, H. (2010). LETOR: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374.

Radlinski, F. and Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–248.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1995). Okapi at TREC-3. *Nist Special Publication Sp*, 109:109.

Sculley, D. (2009). Large Scale Learning to Rank. *NIPS 2009 Workshop on Advances in Ranking*, pages 1–6.

Tax, N., Bockting, S., and Hiemstra, D. (2015). A Cross-Benchmark Comparison of 87 Learning to Rank Methods. In *Information Processing and Management*.

Turnbull, D. and Berryman, J. (2016). *Relevant Search*. Manning Publications Co., New York.

Wang, D., Chen, W., Wang, G., Zhang, Y., and Hu, B. (2010). Explore click models for search ranking. *International Conference on Information and Knowledge Management, Proceedings*, pages 1417–1420.

Wang, X. and Bendersky, M. (2018). Google AI Blog: TF-Ranking: A Scalable TensorFlow Library for Learning-to-Rank.

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., and Najork, M. (2018). Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, New York, NY, USA. ACM.

White, R. and Roth, R. (2008). *Exploratory Search*. Morgan & Claypool Publishers.

Xu, J., Wei, Z., Xia, L., Lan, Y., Yin, D., Cheng, X., and Wen, J.-R. R. (2020). Reinforcement Learning to Rank with Pairwise Policy Gradient. In *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, number 20 in 1, page 10, New York, NY, USA. ACM.

Zhang, Y., Wu, H., Liu, H., Tong, L., and Wang, M. D. (2019). Improve Model Generalization and Robustness to Dataset Bias with Bias-regularized Learning and Domain-guided Augmentation. *arXiv*.