Towards Information Retrieval Evaluation with Reduced and Only Positive Judgements

Diego Mollá
Department of Computing
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au

David Martinez
NICTA and
University of Melbourne
Melbourne, Australia
david.martinez@nicta.com.au

Iman Amini
NICTA and
RMIT
Melbourne, Australia
iman.amini@rmit.edu.au

ABSTRACT

This paper proposes a document distance-based approach to automatically expand the number of available relevance judgements when those are limited and reduced to only positive judgements. This may happen, for example, when the only available judgements are extracted from a list of references in a published clinical systematic review. We show that evaluations based on these expanded relevance judgements are more reliable than those using only the initially available judgements. We also show the impact of such an evaluation approach as the number of initial judgements decreases.

Categories and Subject Descriptors

H.2.4 [**Systems**]: Textual databases; H.3.4 [**Systems and Software**]: Performance evaluation

Keywords

Information Retrieval, Evaluation, Relevance Judgements Expansion

1. INTRODUCTION

There are applications that benefit from an information retrieval (IR) stage, but which do not have enough sample documents for a full assessment of the retrieval quality. Furthermore, the few sample documents available only represent positive relevant documents. For example, within the area of Evidence Based Medicine (EBM), clinical systematic reviews provide the medical doctor with clinical evidence together with a list of relevant documents. We envisage the development of tools that will facilitate the production of such systematic reviews. One of the first stages of such an application consists of an IR step that retrieves all key relevant documents. But the references in a systematic review cover only a small sample of all relevant references [2], and only a fraction of the documents of a systematic review can be retrieved after performing exhaustive searches, mostly due to the fact that there are complex queries and several document repositories [6]. Furthermore, the list of references only indicate relevant documents but there are no lists of non-relevant documents readily available. It is therefore expected that any evaluation metric that is based solely on the references from the systematic review will show unreliable results.

Previous work has shown that by expanding an initial set of document assessments for given queries, one can perform a more accurate automatic evaluation of IR systems. For example, Büttcher et al. [1] used Machine Learning methods trained over a subset of relevance judgements in order to expand the set of relevance judgements. They showed that evaluation results with the expanded set of relevance judgements had better quality than using the source subset of judgements. Quality of the evaluation was measured by ranking a set of IR systems according to the new expanded relevance judgements, and comparing it against the system ordering produced by the original set of judgements. In the clinical domain, Martinez et al. [6] explored the use of reranking methods based on reduced judgements, and found that the use of automatic classifiers would allow to considerably reduce the time required for clinicians to identify a large portion (95%) of the relevant documents. Both these articles reported limitations of the classifiers when the initial number of documents was small. Furthermore, in the scenario that we contemplate, where we rely on the list of references of a systematic review as the set of relevant documents, we do not have information about negative judgements, and therefore a classifier-based approach to expand the set of relevant documents would have to deal with this issue.

More recent work [7] has shown that by relying on documents retrieved frequently by a diverse set of systems, it is possible to build relevance assessments automatically, and achieve high correlation with manually judged data. However this approach has been tested by building on a set of competing runs from different research groups, which is not always available; and this method does not benefit from existing qrels.

We propose to automatically expand the set of relevant documents by adding documents that are reasonably close to the original, reduced set. We show the result of several experiments that test the impact of such automatic expansion. For our experiments, we rely on the OHSUMED test collection [3]. This is a corpus containing clinical queries and assessments, and we focus on the set of 63 queries that was used in the TREC-9 Filtering Track. The OHSUMED

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ADCS '13, December 05 - 06 2013, Brisbane, QLD, Australia
Copyright 2013 ACM 978-1-4503-2524-0/13/12 ...$15.00.

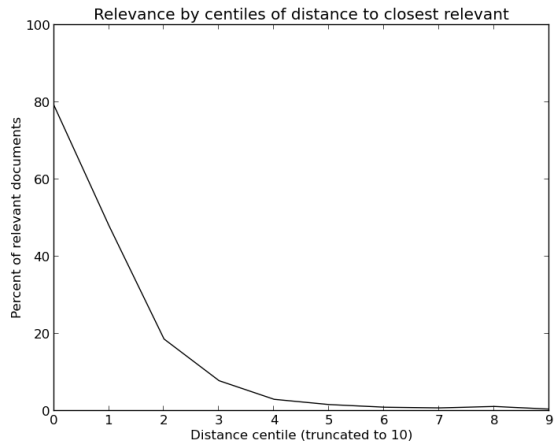

Figure 1: Distance versus relevance in the OHSUMED test corpus.

queries were generated to address actual information needs for clinicians, and the assessed documents were retrieved in two iterations, by relying on the MEDLINE search interface[1] and the SMART retrieval system respectively. The retrieved documents were judged by a separate group of domain experts to the group performing the search. As document collection we rely on the 1988-91 subset of MEDLINE that was released as test data for the TREC-9 challenge, which contains 293,856 documents. For evaluation we apply a variety of IR systems implemented in the Terrier open source package [4].

2. DISTANCE VERSUS RELEVANCE

The rationale of our work is related to the so-called cluster hypothesis, that is, the assumption that "documents that are in the same cluster behave similarly with respect to relevance to information needs" [5]. The cluster hypothesis has been used to improve the results of information retrieval and classification tasks. In contrast, we are not concerned about improving the IR results. Instead, we want to improve the effectiveness of IR evaluation. But this slightly reworded version of the cluster hypothesis may apply: documents *that are similar enough* will behave similarly with respect to relevance to information needs. The question is, how similar must these documents be?

We first examined the impact of similarity between documents with regards to their relevance. For every document associated to any qrel from the OHSUMED test set (3,121 documents), we computed the distance between the document and the closest qrel (other than the document itself) within each query. The resulting (document,question) pairs were sorted by distance and binned into centiles such that the first centile is formed by the top 1% pairs, and so on. Then, within each centile we computed the percentage of relevant documents. Figure 1 shows the result.

The figure shows a clear relation between distance and relevance. 78% of documents in the first centile are relevant, and the number quickly degrades. The figure has been truncated to the top 10 centiles since virtually none of the documents from the 10th centile onwards are relevant.

For these experiments we used as the distance metric $1 - \text{cosine similarity}$. The vector representations of the documents were formed by obtaining the tfidf values of all words lowercased and with stop words removed, and then taking the top 200 components after performing Principal Component Analysis (PCA).[2] PCA was used as a means to compress the vector space. Using tfidf features without the subsequent PCA stage produced a slightly less marked relation and at the expense of much longer processing times.

[1]http://www.ncbi.nlm.nih.gov/pubmed

BB2	BM25	DFR_BM25	DLH
DPH	DFRee	Hiemstra_LM	DLH13
IFB2	In_expB2	In_expC2	InL2
LemurTF_IDF	LGD	PL2	TF_IDF

Table 1: List of 16 runs from the terrier package

3. EVALUATION METRICS

The results described in Section 2 show that distance between a document and a known relevant document may be a good indicator of document relevance. We therefore studied the impact of using document distance as a means to generate new relevance sets which we call pseudo-qrels.

In the following experiments we used several IR systems as described below. We evaluated the performance of each system according to these sets of relevance judgements:

1. Original set of qrels.
2. A subset of qrels. This subset is a baseline that models the situation where the number of qrels is limited.
3. The same subset of qrels, expanded with the pseudo-qrels. These pseudo-qrels are produced based on distance metrics as described below.

3.1 Information Retrieval Baselines

In the absence of the official set of runs from the TREC filtering track participants, we resorted to building our own systems using the open source Terrier 3.5 package. We built 16 baselines by choosing 16 different ranking algorithms and used them with their default settings to build the runs.

Terrier offers a range of Divergence from Random (DFR) models which are instantiated by three components of the framework: selecting a basic randomness model, applying the first normalisation, and normalising the term frequencies. We stopped and stemmed all the 63 test queries and the collection, and used the Porter stemmer as the default stemming algorithm. Table 1 is the list of ranking models corresponding with the baselines used for our experiments.

3.2 Pseudo-qrels for Evaluation

The pseudo-qrels of a query are generated by selecting those that are closest to some qrel within the query, using the $1 - \text{cosine}$ distance metric described in Section 2:

1. For every query q:
   (a) For every document d in the pool of available documents:

[2]These experiments were carried out in Python and the scikit-learn library.

Train Qrels	Test Qrels	Precision	Recall	F-score
10%	90%	0.360	0.100	0.157
20%	80%	0.345	0.118	0.176
30%	70%	0.290	0.112	0.161
40%	60%	0.282	0.125	0.173
50%	50%	0.244	0.123	0.164

Table 2: Retrieved pseudo-qrels evaluated against the original relevance set.

i. Record the minimum distance between d and the set of qrels within q (except d).

2. Sort the resulting triples (distance,d,q) in ascending order and select the top K.

3. Add the selected documents d to the corresponding q. These are the pseudo-qrels.

For these experiments, the pool of available documents was generated by taking the top N documents retrieved by each query for all of the IR systems that we used. We varied the percentage of available qrels in our experiments, always making sure that each query had at least one qrel.

Note that the above algorithm selects the candidate pseudo-qrels using a threshold that is global to all queries. This means that some queries may receive more pseudo-qrels than others, and in extreme cases only a few queries will receive pseudo-qrels. We thought that this is desirable, since the experiment in Section 2 shows such a strong impact of document distance in the relevance of the document. If a query only has documents that are relatively far from known qrels, we better not add them as pseudo-qrels.

The approach described above can be seen as a simple one-rule classifier based on distance. We resorted to such a simple classifier instead of a more sophisticated classifier such as the SVM classifier used by Büttcher et al. [1] because of the scarcity of data and lack of negative judgements in our scenario. If we were to train an SVM classifier we would need to find a means to reduce overfitting.

For a first estimation of the quality of the retrieved qrels, we evaluated our method in the manner of a text classification system, by relying on different splits of the original qrels, and measuring the F-score value for the detection of relevant documents for each query. For this experiment we use different partitions as "training" data (which is used for the documents in the collection to compare against) and "test" data (which is used for evaluation).

In order to retrieve pseudo-qrels, we set $N = 100$ (we retrieved the top-100 documents for each of our IR systems), and $K = 0.2\%$ (we considered as relevant the top 0.2% of the most similar documents). The evaluation of the pseudo-qrels is given in Table 2, when using up to 50% of the qrels as training data. We can see that overall the performances are low, specially in recall, but Büttcher et al. [1] found that low F-scores can still lead to large improvements when measuring the correlation between manual and semi-automatic relevance judgements. The results also illustrate that when using only 20% of relevant documents, we achieve the highest F-score, and more than a third of the retrieved documents are relevant.

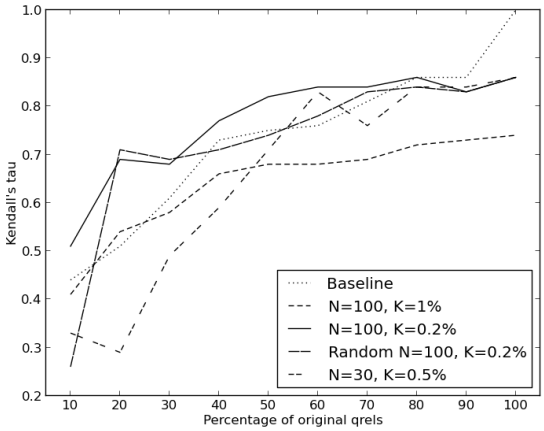

Figure 2: Kendall's tau of system orderings using MAP. The baseline uses percentages of the original qrels. The other evaluations use percentages of the qrels plus the pseudo-qrels for several choices of N (number of documents chosen per query) and K (percentage of top documents selected as qrels).

3.3 Correlation for ranking IR systems

Figure 2 shows Kendall's tau between the ranking of the IR systems when evaluated using 1) a baseline consisting of original qrels, and 2) varying percentages of qrels extended with the computed pseudo-qrels. The evaluation metric was MAP. The figure presents the results for varying values of N (the number of documents taken from each query in each IR system), and K (the percentage of top documents selected as pseudo-qrels).

The baseline shown in the figure uses the qrels without the pseudo-qrels and it reflects the quality of the evaluation when using the available data. We can observe, as expected, that larger percentages of qrels lead to better correlation figures. The other curves show the evaluation quality when the qrels are expanded with pseudo-qrels.

The figure shows that different choices of values of N and K affect the quality of the evaluation. When we choose relatively few documents ($N = 30$) to form the pool of available documents, the results do not improve on the baseline. This is presumably due to the lack of enough documents to gather useful statistics. When we choose a larger number of documents ($N = 100$), then a wise threshold K may lead to improvements. In our experiments, choosing a relatively small percentage of documents from the pool ($K = 0.2\%$) leads to results above the baseline, but choosing a larger percentage ($K = 1\%$) leads to a decline of results. This is in line with the analysis shown in Figure 1, which indicates that the percentage of relevant documents decreases steeply as we increase the distance. Therefore a threshold which is too relaxed may introduce too much noise. With a choice of N=100 and K=0.2%, small percentages of qrels lead to a comparatively greater improvement over the baseline. These results are very encouraging and support the idea of using distance metrics to compensate for the lack of available relevance judgements and the lack of negative relevance judgements.

When selecting the qrels, all the results described above

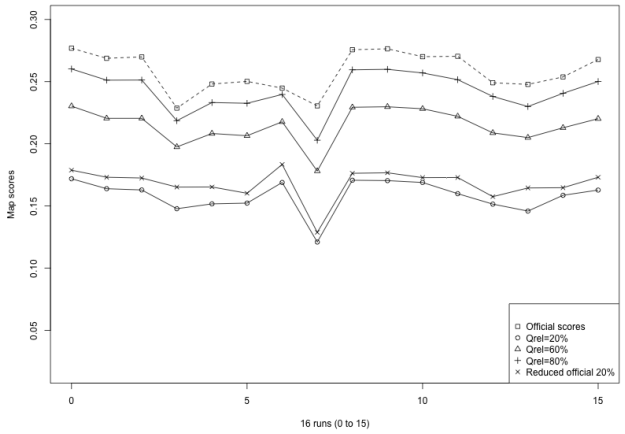

Figure 3: Official map scores using official qrel versus limited true qrels combined with the pseudo-qrels for N = 100 and K = 0.2

used the top qrels (those appearing first in the list of qrels). Figure 2 also includes the results when using a random selection of qrels. It shows wide changes for small percentages of qrels, and it tends to agree with the baseline for larger percentages. This probably means that the choice of qrels really matters, and documents from the top qrels may be quality relevant documents. In future work we will study the impact of the selection of qrels further.

Figure 3 shows the system map scores using the official qrel combined with the pseudo-qrels for three varying sizes of limited positive qrels. As a reference we also show the curve when using 20% of qrels only. It can be seen that the scores generated by the pseudo-qrels range in the vicinity of the official map scores for qrel = 80%, while the results with the lower percentage of true qrels tend to be underestimated. However, the ordering of the runs which was our ultimate goal, remains stable. When using 20% of qrels, the curves with and without pseudo-qrels look similar, but lead to different rankings, as the Kendall's tau scores in Figure 2 illustrate.

4. CONCLUSIONS

We have shown promising results towards the use of a simple distance-based approach to expand a set of relevance judgements. The results are particularly encouraging when the number of available relevance judgements is very limited, and works when there are only positive judgements.

These results suggest the use of distance-metrics extensions of relevance judgements as a quick and cheap evaluation during the development stage of information retrieval systems when there are few and only positive relevance judgements. It can therefore be applied for the development of IR systems that search for relevant clinical studies, even when the set of known available relevant documents is just the list of references of a sample clinical systematic review.

Further work includes a more comprehensive study of the thresholds that lead to the best evaluation setting. It is also desirable to determine how well these findings carry to other domains. Also, given that the measure of quality used in this study is based on the correlation of rankings with an automated evaluation metric, it is desirable to extend this study with real human judgements for a more precise characterisation of the possibilities of this approach.

We have used a very simple distance metric in this study. It will be interesting to explore the impact of additional distance metrics that may use domain knowledge or more sophisticated linguistic information.

5. ACKNOWLEDGMENTS

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

6. REFERENCES

[1] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 63, New York, New York, USA, 2007.

[2] K. Dickersin, R. Scherer, and C. Lefebvre. Identifying Relevant Studies for Systematic Reviews. *BMJ (Clinical research ed.)*, 309(6964):1286–91, 1994.

[3] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. pages 192–201, 1994.

[4] C. Macdonald, R. McCreadie, R. Santos, and I. Ounis. From Puppy to Maturity: Experiences in Developing Terrier. *Open Source Information Retrieval*, page 60, 2012.

[5] C. D. Manning. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (MA), 2008.

[6] D. Martinez, S. Karimi, L. Cavedon, and T. Baldwin. Facilitating Biomedical Systematic Reviews Using Ranked Text Retrieval and Classification. In *Australasian Document Computing Symposium (ADCS)*, pages 53–60, Hobart, Australia, 2008.

[7] T. Sakai and C.-y. Lin. Ranking Retrieval Systems without Relevance Assessments - Revisited. In *The Third International Workshop on Evaluating Information Access (EVIA)*, pages 25–33, 2010.