Towards automatic generation of relevance judgments for a test collection

Mireille Makary / Michael Oakes
RIILP
University of Wolverhampton
Wolverhampton, UK
m.makary@wlv.ac.uk / michael.oakes@wlv.ac.uk

Fadi Yamout
Computer Science Department
Lebanese International University
Beirut, Lebanon
fadi.yamout@liu.edu.lb

Towards automatic generation of relevance judgments for a test collection

Mireille Makary / Michael Oakes
RIILP
University of Wolverhampton
Wolverhampton, UK
m.makary@wlv.ac.uk / michael.oakes@wlv.ac.uk

Fadi Yamout
Computer Science Department
Lebanese International University
Beirut, Lebanon
fadi.yamout@liu.edu.lb

***Abstract*—this paper represents a new technique for building a relevance judgment list for information retrieval test collections without any human intervention. It is based on the number of occurrences of the documents in runs retrieved from several information retrieval systems and a distance based measure between the documents. The effectiveness of the technique is evaluated by computing the correlation between the ranking of the TREC systems using the original relevance judgment list (qrels) built by human assessors and the ranking obtained by using the newly generated qrels.**

Keywords—Evaluation; qrels; document distance; occurrences; test collections; relevance judgments

I. INTRODUCTION

Information retrieval is the process of retrieving relevant information to satisfy the user's need which was expressed by formulating a query and submitting it to an information retrieval system. Given different systems, how can we determine which one performs best? When we implement new retrieval algorithms, how can we test their performance compared to other existing algorithms? We use test collections for this purpose. A test collection is a set of documents, a set of manually constructed topics, and a relevance judgments list (also called query based relevance sets, qrels) which is built by human assessors. This relevance judgment list shows the topic number, the document id and the document's binary relevance to the topic, where "1" indicates relevance and "0" non-relevance.

This is known as the Cranfield paradigm, which was first started by Cleverdon in 1957[1]. It involves manual indexing for the documents, and assessing all documents from a database for relevance with respect to a finite set of topics. The Text REtrieval Conference (TREC) organized annually by NIST provides such a framework to allow larger-scale evaluations for text retrieval. TREC provide test collections, each with a relevance judgment list built by human assessors based on a pooling technique (Spärck Jones and van Rijsbergen) [2]. Each TREC test collection has 50 topics and a set of documents. All participating research groups are given these documents. Each group uses the topics provided and retrieves a ranked set of documents using their information retrieval system. They then submit their runs back to NIST. The researchers at NIST will then form a pool of documents of depth 100 for each topic, by collecting the top 100 documents from each run. Duplicate documents are then removed. Each document in the resulting pool is then judged by human assessor to determine its relevance. This forms the relevance judgment list or the query-based relevance sets (qrels). Any document not found in the pool is considered to be non-relevant. Building the qrels is a major task and consumes a lot of time, resources and money. It becomes practically infeasible when the test collection is huge and contains millions of documents. This is why various researchers have worked to automate the generation of the qrels or build them with minimal human intervention. The Cranfield paradigm is still widely used mostly for academic and partially commercial system evaluation. It is also still important in traditional ad hoc retrieval both in specific tasks and for certain web queries, but Harman has spoken on possible future modifications [16].

In this paper, we devise a new methodology to build the set of qrels without any human intervention. The structure of the remainder of this paper is as follows: In section 2 we review the previous work done in this field. In section 3 we describe the experimental design for a new system of producing qrels completely automatically, and in section 4 we give the results of experiments which show that our new system outperforms the earlier systems which inspired it. In section 5 we conclude with some ideas for future work.

II. RELATED WORK

Zobel [3] explained how it is possible to use the top retrieved documents to predict with some accuracy how many relevant documents can still be found further down the ranking, but this methodology was not tested. Interactive searching and judging proposed by Cormack et al [4] is an interactive search system that selects the documents to be judged. It uses Boolean query construction and ranks documents based on their lengths and the number of passages that satisfy the query. Search terms will be highlighted to help assessors in judging the documents. Searchers by this technique try to find as many relevant documents as possible for each of the topics included. The move-to-front (MTF) technique [4] directly improves the

TREC baseline pooling method since it selects different numbers of documents depending on the system performance. As opposed to TREC pooling, it examines the documents in order of their estimated likelihood of relevance. Soboroff et al. [5] proposed that manual relevance assessments could be replaced with random sampling from pooled documents. From the previous TREC results, they developed a model of how relevant documents occur in a pool. This was achieved by computing the average number of relevant documents found per topic in the pool, and the standard deviation. However, this information is not available in practice for systems not trained on TREC data. A related method was suggested by Aslam and Savell [6] who devised a measure for quantifying the similarity of the retrieval systems by assessing the similarity of their retrieval results. The use of this new measure evaluated system performance instead of system popularity, so that novel systems which produced very different sets of qrels to the others were not penalized. Nuray and Can [7] generated the relevance judgments using heuristics. They replicated the imperfect web environment and modified the original relevance judgment to suit the web situation. They used the pooling technique described earlier and then ranked the documents based on the similarity score of the vector space model. Carterette et al [8] linked the evaluation of an IR system using the Average Precision (AP) to the construction of test collections. After showing that AP is normally distributed over possible sets of relevance judgments, a degree of confidence in AP was estimated. This new way of looking at the evaluation metric led to a natural algorithm for selecting documents to judge. Efron's method used query aspects [9], where each TREC topic was represented using manual and automatically generated "aspects". The same information need might be represented by different aspects. Each manually derived aspect was considered as a query and the union of the top 100 documents retrieved for each topic was considered to be the set of "pseudo-qrels" or "aspect qrels". Other techniques were an improvement to the pooling technique. In their experiments to build a test collection, Sanderson and Joho [10] obtained results which led them to conclude that it is possible to create a set of relevance judgment lists (RJL) from the run of a single effective IR system. However, their results do not provide as high a quality set of qrels as those formed using a combination of system pooling and query pooling as used in TREC.

The power of constructing a set of information "nuggets" extracted from documents to build test collections was shown by Pavlu et al [11]. A nugget is an atomic unit of relevant information. It is a sentence or a paragraph that holds a relevant piece of information which leads to the document being judged as relevant. Rajput et al. [12] used an "Active Learning" principle to find more relevant documents once relevant nuggets are extracted, because a relevant document infers relevant information and relevant information leads to finding more relevant documents.

III. EXPERIMENTAL DESIGN

The technique used in this paper is inspired by both Rajagopal and Mollá techniques [14] [13] which are described in the following sections.

A. Rajagopal's technique

Rajagopal[14] used two independent approaches to build pseudo relevance judgements: one which is completely automated does not require any human intervention and is based on a "cutoff percentage" of the number of documents to mark as relevant or non-relevant. The second is called "exact count" and it requires previous knowledge of the number of documents judged relevant by the human assessor. The results they obtained showed that the approach based on cutoff percentage gave better Kendall's tau and Pearson correlation values between system rankings based on humanly-annotated qrels and machine-genrerated qrels. Since in this paper we are interested in completely automating the process of building relevance judgment lists, and the aim is to prove that we can suggest a new technique that can provide better correlation values, we will describe and compare our results against the "cutoff percentage" technique only. Rajagopal's technique used the number of occurrences of a document in each system run to determine its relevancy, whether it is relevant or non-relevant to a topic. The hypothesis made initially states the following: the higher the number of occurrences of a document in the pool of documents found relevant by a range of systems, the higher is the probability of this document being relevant. In their experiment, a variation of the TREC pooling technique was presented, since pseudo relevance judgments are built without any human assessors' involvement. Cutoff percentages (>50% and >35%) of documents occurrences were studied. A pool depth of 100 was used. The steps followed for TREC-8 were: (1) Get the runs from all the systems, (2) pool with depth K (here K =100), (3) calculate the number of occurrences per document per topic, (4) order by the number of occurrences of documents per topic in descending order, (5) calculate the % values of these occurrences, therefore, for a total of 129 systems, if doc1 occurred in 10 systems, the percentage value is about 7%, (6) set document relevancy based on the cutoff percentage. So if for topic 1 doc34 had a percentage value of 64%,, it will be considered relevant otherwise depending on the cutoff percentage chosen (50% or 35%) if it is below this cutoff, it will be considered non-relevant (7) Calculate MAP for all systems, rank them and compute the correlation. The results reported by Rajagopal are shown in Table 1:

TABLE I.

TREC-8 (129 Systems)	**Kendall's tau**	**Pearson**	**Harmonic Mean**
cutoff >50%	0.506	0.739	0.600
cutoff >35%	0.515	0.736	0.605

Table1: *Kendall's tau and Pearson correlation for MAP values for depth 100 using different cutoff percentage for TREC-8*

A question that extends from the above experiments: does increasing the cutoff percentage provide better results? What will be the correlation obtained for cutoff percentages greater than 50%, such as 60% and 80%? The reason behind increasing the cutoff percentage is to minimize the error margin when judging documents as relevant and this is needed to expand the positive judgments using Mollá's technique for measuring the similarity between the documents. A description of the distance based measure used to compare documents is described below.

B. Mollá's Technique

Mollá [13] used a distance based measure to expand positive judgments only. The distance measure was based on the cosine similarity measure [15] between two document vectors. The distance measure is defined by:

Distance_measure = 1 – cosine measure (1)

The hypothesis was that relevant documents are at a close distance to each other, so they form a cluster. To prove it, he used different Terrier weighting models as surrogates for different retrieval systems. He measured the distance between some known qrels and the document retrieved. If it was less than a certain threshold, the document was considered relevant. He then evaluated the system rankings by using the original qrels, a subset of the qrels and then the same subset selected in the previous experiment with the expanded list of documents automatically judged relevant added. However, his method requires knowing a set of relevant documents a priori and then expanding only positive judgments.

C. New Technique

The new technique used in this paper does not require any human intervention and has no prior knowledge of the test collection's original qrels. We used the TREC-8 test collection in our experiments and we tested using the 129 TREC systems. We followed first the same steps done by Rajagopal only now we chose different cutoff percentages (>=60% and >=80%). We select the documents that were retrieved by more than 60% or 80% of the systems. The purpose of increasing the cutoff percentage was to ensure having a high probability set of relevant documents. Because the set returned by a cutoff percentage of 80% contained more relevant documents, we used this set (called **(S)**) to find more relevant documents in the pool by using the similarity measure similarity in equation (1). For each document $\mathbf{(d_i)}$ in the pool of depth 100 created by all 129 systems, we measured the distance between $\mathbf{(d_i)}$ and each document in the cutoff set **(S)** formed for a topic **i.** We selected the closest pair of documents. Only when the distance between each pair was less than a threshold **(ε)** determined empirically, the document was marked relevant otherwise it was marked non-relevant. We evaluated our technique by computing the MAP values for each of the TREC systems and comparing the different rankings obtained when using the original qrels and the newly generated ones. For different values of **(ε)**: 0.5, 0.4, 0.3, 0.28, 0.26, 0.2 and 0.15, the Pearson correlation showed better value for **ε** =0.2 while the Kendall's tau is better for **ε** = 0.4. The correlation values for each experiment conducted are given in the next section.

IV. Results and Discussion

Here we describe the evaluation process of the new technique. We compute the MAP value for each of the TREC systems using the original set of qrels that were built by human assessors and rank those systems. Then we compute the MAP based on the newly generated qrels and we rank the TREC systems. We measure the correlation between the two rankings by computing the Pearson and Kendall's tau coefficients. For the first experiment that follows Rajagopal's cutoff percentage technique, the results from using cutoff percentages of 60% and 80% are shown below in table 2:

TABLE II.

TREC-8 (129 Systems)	**Kendall's tau**	**Pearson**	**Harmonic Mean**
cutoff >=60%	0.507	0.748	0.604
cutoff >=80%	0.489	0.766	0.597

Table 2: *Kendall's tau and Pearson coefficient for TREC-8 experiments using TREC systems based on cutoff percentages*

A cutoff percentage of 80% provides the best correlation value even though the Kendall's tau coefficient is less by 2.6% than the 35% cutoff tested by Rajagopal.

When using different cutoff percentages, we computed the percentage of actual relevant documents retrieved because in reality not all documents retrieved in the cutoff set were judged relevant by human assessors. Table 3 shows that with a cutoff percentage of 80%, almost 24% of the documents considered relevant were actually judged relevant by human assessors and therefore we used this set (S) in the remainder of the experiment to expand the first set of qrels generated and judge more documents as relevant using the distance measure in equation (1).

TABLE III.

For cutoff >=50, percentage of actual relevant docs is:	**For cutoff >=60, percentage of actual relevant docs is:**	**For cutoff >=80, percentage of actual relevant docs is:**
11.9 %	14.4%	23.9%

Table 3: *Percentage of actual relevant documents found in the set automatically judged for different cutoff percentages*

Relevant documents are at a close distance to each other, and in a sense they form a cluster [13]. Now that we have considered the documents retrieved by 80% of the systems as relevant, we tried to judge more documents in the pool of depth 100 as relevant based on the distance measure in (1). For each document retrieved in the pool, we computed the distance between this document and the set of documents that

belong to the cutoff set (S). For example, for topic 401, we have 5 documents that were retrieved by more than 80% of the systems and therefore marked as relevant: D={d1,d2,d3,d4,d5}, so for each remaining document (d) in the pool that was retrieved for topic 401, we computed the distance between (d) and each document in D. The pair of documents where the distance between the (d) and (d4) is the smallest is selected. Now to judge whether (d) is relevant or not, we check the distance value obtained. If it is less than a distance threshold value ε (determined empirically), (d) will be marked as relevant otherwise it will be marked as non-relevant. This process is repeated for each document in the pool retrieved for a topic and for each of the 50 topics. At the end, we will have a new set of qrels that was automatically built without any manual intervention.

We tried different values for the distance threshold (ε) and we computed the Kendall's tau and Pearson coefficients for evaluation (table 4).

TABLE IV.

Threshold (ε)	Kendall's tau	Pearson	Harmonic Mean
0.5	0.4451	0.7017	0.5446
0.4	0.5033	0.7654	0.6072
0.3	0.5032	0.7804	0.6118
0.2	0.4879	0.7814	0.6007
0.15	0.4809	0.7786	0.5945

Table 4: *Kendall's tau and Pearson coefficients for different values of the distance measure threshold*

The results show that the best Kendall's tau value is obtained for ε=0.4 while the best Pearson value is for ε=0.2. But as an overall comparison between the results using the harmonic mean of the two measures, the best value is achieved for ε=0.3. In all cases, the Pearson coefficient shows better results than obtained when using different cutoff percentages only.

We divided the TREC systems into three subsections based on the retrieval effectiveness values, the MAP value: the top third of the systems are considered to be good performing systems, the middle third are the moderate performing systems and the bottom third are the low performing systems. Grouping the systems into different groups is done to identify if our approaches perform better for a specific subsection of systems than the other. We then computed the Kendall's tau and Pearson values for each subsection based on the results achieved by Rajagopal's cutoff >50% approach, our cutoff >=80% and cutoff >=80% with ε=0.3 approaches. The results were very similar. The correlation between the low performing systems seems to be the best. The automatically generated qrels using a cutoff >=80% are most effective in discriminating among poorly performing systems. As for the other two subsections, the correlation falls below 0.5 (tables 5 and 6). The negative value obtained for good and moderately performing systems indicates that when the rank of one system increases in the original rank, it decreases in the rank obtained by the newly generated qrels or vice versa. This could be resulting from the fact that some systems are contributing to the new set of qrels automatically built based on the cutoff percentage or distance based measure while it was not contributing in forming the original qrels. Also in TREC when a document is retrieved from a noncontributing system, it is marked as non-relevant, but in our case we might have marked it as relevant because the number of occurrences is above the cutoff percentage defined.

TABLE V.

Methods	Good Performing Systems	Moderately Performing Systems	Low Performing Systems
Cutoff >50% (Rajagopal's)	-0.2313	0.3842	0.7799
Cutoff >=80%	-0.2546	**0.3953**	**0.7928**
Cutoff >=80% and ε=0.3	**-0.2174**	0.3324	0.7773

Table 5: *Kendall's tau correlation for 3 subsections for depth 100 using different cutoff percentages and the distance based approach for TREC-8*

TABLE VI.

Methods	Good Performing Systems	Moderately Performing Systems	Low Performing Systems
Cutoff >50% (Rajagopal's)	-0.8111	**0.5919**	0.9169
Cutoff >=80%	**-0.8088**	0.5681	**0.9483**
Cutoff >=80% and ε=0.3	-0.8128	0.5066	0.9435

Table 6: *Pearson correlation for 3 subsections for depth 100 using different cutoff percentages and the distance based approach for TREC-8*

As an overall value, we compute the harmonic means for Kendall's tau and Pearson correlations for each subsection of the systems and the values obtained by our proposed cutoff >=80% approach and the one that expands the positive judgments based on the distance measure seem to provide better values.

TABLE VII.

Methods	Good Performing Systems	Moderately Performing Systems	Low Performing Systems
Cutoff >50% (Rajagopal's)	-0.3599	0.4659	0.8428
Cutoff >=80%	-0.3872	**0.4662**	**0.8636**
Cutoff >=80% and ε=0.3	**-0.3430**	0.4014	**0.8523**

Table 7: *Harmonic means for 3 subsections for depth 100 using different cutoff percentages and the distance based approach for TREC-8*

To perform an intrinsic evaluation for the qrels automatically generated, we compute the precision and recall measures at different ranks (@5, @10, and @20… @ 100, @ 20 ... @ 1000). The formula used for the precision metric is shown in (2)

$$\text{Precision} = d_{AH} / d_A \qquad (2)$$

Where $\mathbf{d_{AH}}$ is the number of documents judged relevant automatically by new technique and human judge and $\mathbf{d_A}$ is the

number of documents judged relevant automatically by new technique.
As for the recall metric, the formula used is described in (3).

$$\text{Recall} = d_{AH} / d_H \qquad (3)$$

Where $\mathbf{d_{AH}}$ is also the number of documents judged relevant automatically by new technique and human judge and $\mathbf{d_H}$ is the number of documents judged relevant by human assessors.

We also computed the precision and recall for the qrels generated by Rajagopal's technique for a cutoff percentage >50%. Figure 1 plots the precision values at different ranks for Rajagopal's technique using the 50% cutoff percentage and the new technique using a distance threshold of 0.2. As it can be seen our technique outperforms the values obtained by Rajagopal's at almost every rank except at rank 5 where the precision is really close (0.1 – Rajagopal and 0.08 using the new technique). For the recall, the cutoff of 50% scores better recall values than our technique using a distance threshold of 0.2. But if we increase the distance threshold to 0.5, our method can achieve similar or even better scores at some ranks as the plot in Figure 2 shows.

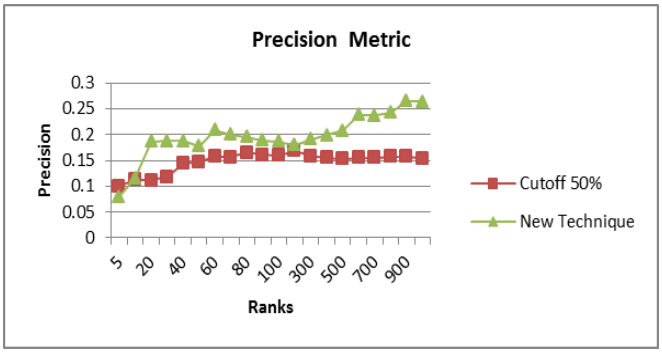

Fig. 1. *Precision metric at different ranks for both techniques: the one using a cutoff percentage 50 and the new proposed technique using a distance threshold of 0.2*

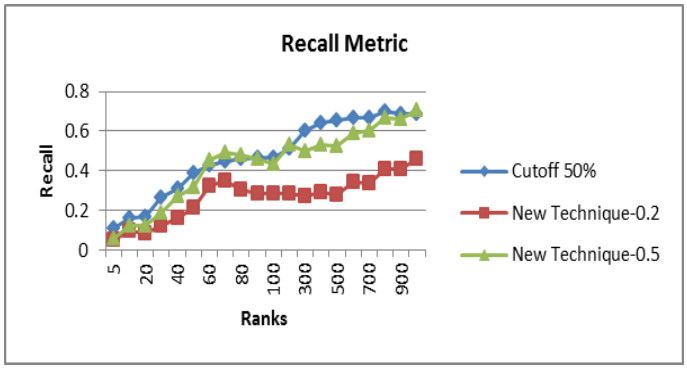

Fig. 2. *Recall metric at different ranks for techniques: the one using a cutoff percentage 50 and the new proposed technique using a distance threshold of 0.2 and of 0.5.*

In conclusion, the technique we propose in this paper can provide a set of qrels which correlates better (compared with the earlier systems) with the ones formed by humans than using a cutoff percentage based technique and when performing both the intrinsic evaluation (recall and precision of the discovered document sets) and the extrinsic (ability to rank systems compared with the original TREC documents), we achieve values for different distance threshold. Therefore, this method allows us to reduce cost and time when building test collections for system evaluation.

V. CONCLUSION

In this paper, we used a combination of pooling retrieved documents and clustering based on the distance between them in the vector space model to build a set of relevance judgments or qrels for a test collection without any human intervention. The approach we use allows expanding the set of qrels based on a distance measure between the documents. The technique is independent of the test collection type so this might guide us towards new experiments in which we can built a set of qrels for non-TREC test collections and it will be interesting to study its use with non-English test collections.

REFERENCES

[1] Cleverdon C. The cranfield tests on index language devices. Aslib Proceedings, Volume 19, pages 173– 192, 1967.

[2] Spärck Jones K. and van Rijsbergen C.J. Information retrieval test collections, Journal of Documentation, 32, 59-75, 1976.

[3] Zobel J. How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 307-314, 1998.

[4] Cormack G.V., Palmer C.R. and Clarke C.L.A. Efficient Construction of Large Test Collections, in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 282-289, 1998.

[5] Soboroff I., Nicholas C., and Cahan P. Ranking retrieval systems without relevance judgments, In Proceedings of ACM SIGIR 2001, pages 66–73, 2001.

[6] Aslam J. A. and Savell R. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In Proceedings of ACM SIGIR 2003, pages 361–362, 2003.

[7] Nuray R. and Can F. Automatic ranking of information retrieval systems using data fusion, Information Processing and Management, 42:595–614, 2006.

[8] Carterette B., Allan J., Sitaraman R.: Minimal test collections for retrieval evaluation. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, WA, ACM Press 268-275, 2006.

[9] Efron M.: Using multiple query aspects to build test collections without human relevance judgements, SIGIR, 2009.

[10] Sanderson M., Joho H.: Forming test collections with no system pooling. In: SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM 33-40, 2004.

[11] Pavlu V., Rajput S., Golbus P. B., and Aslam J. A. IR system evaluation using nugget-based test collections, WSDM '12, 2012.

[12] Rajput S., Ekstrand-Abueg M., Pavlu V., Aslam J. Constructing Test Collections by Inferring Document Relevance via Extracted Relevant Information, CIKM '12 Proceedings of the 21st ACM international conference on Information and knowledge management Pages 145-154.

[13] Mollá D., Martinez D., Amini I. Towards information retrieval evaluation with reduced and only positive judgments, ADCS '13 Proceedings of the 18th Australasian Document Computing Symposium, Pages 109-112, 2013.

[14] Rajagopal P., Ravana S.D., and Ismail M.A. Relevance judgments exclusive of human assessors in large scale information retrieval evaluation experimentation, 2014.

[15] Salton G. and McGill M. J. (1983) “Introduction to Modern Information Retrieval”. McGraw Hill, New York, 1983.

[16] Harman D. Is the Cranfield Paradigm Outdated? A keynote talk in SIGIR’10 (2010).