International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

BUILDING A RELEVANCE JUDGMENT LIST WITH MINIMAL HUMAN INTERVENTION

Mireille Makary, University of Wolverhampton ; Fadi Yamout,
Lebanese International University; Michael Oakes University of Wolverhampton

Abstract

This paper describes a new technique for building a relevance judgment list (qrels) for TREC test collections with minimal human intervention. We run twelve different Terrier weighting models using TREC topics. For each topic, we collect the common set of documents which were retrieved by all weighting models. Using the Keyphrase Extraction Algorithm (KEA) implemented as a plugin to GATE, we extract key phrases from each of the selected documents. Then, we assign a score to each key phrase based on the number of terms it shares with the original TREC topic. The key phrases with the highest scores become the queries for a second search, this time using the Terrier BM25 weighting model. The union of the documents retrieved forms the set of qrels for the original TREC query. We evaluate the relevance judgment list (qrels) obtained by ranking the twelve weighting models provided by Terrier, used for retrieval (surrogates for different retrieval systems) both with original TREC queries and with the qrels derived by our method, and finding the correlation between these rankings.

Introduction

An Information Retrieval System can be evaluated using test collections based on the Cranfield model [1]. A test collection consists of a corpus of documents, a set of queries and a relevance judgment list (qrels), which is a list containing documents relevant to each query. Building these qrels has always been a time and effort consuming task. It requires human assessors to judge each document and determine whether it is relevant or not to the query or topic in question. This judgment becomes practically infeasible in large scale collections, where the corpus has millions of documents in it. In this paper, we devise a new approach for automatically generating the set of qrels with minimal human intervention.

Related Work

A means of producing a relevance judgment list (qrels) without exhaustive searching was proposed by Spärck Jones and Van Rijsbergen. In [2, 3] the building of an ideal test collection was described. The use of pooling was advocated as a means of efficiently locating relevant documents within a large test collection. For each query, merging the output of diverse searches formed a pool. It was assumed that nearly all the relevant documents would be found in the pool. A random sample of the document pool would then be manually assessed for relevance. The assessed documents form, thereby, the qrels.

Recent years have seen increased interest in methods for evaluating IR systems without human intervention. This type of "judgment-free" evaluation has not yet been fully accomplished.

A number of methods have been devised for building the set of qrels. Soboroff et al. [4] proposed that manual relevance assessments could be replaced with random sampling from pooled documents. From the previous TREC results, they used a model for 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. They showed that documents returned by multiple runs enhance system ranking accuracy. A related method was suggested by Aslam and Savell [5] 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. Efron's method used query aspects [6], where each TREC topic was represented using manual and automatically generated "aspects". Consider TREC topic 402 that has "behavioral genetics" as its title. The same information need might be represented by different aspects such as "behavioral disorders" or "genetics addictions". 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 obtained results which led them to conclude that it is possible to create a set of relevance judgement lists (qrels) from the run of a single effective IR system. However, their results do not provide a high quality set of qrels as those formed using a combination of system pooling and query pooling (as used in TREC) [7].

The power of constructing a set of information "nuggets" extracted from documents to build test collections was shown by Pavlu et al [8]. 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

ISSN:2319-7900

being judged as relevant. For TREC-8 topic 401, titled "foreign minorities, Germany", an example of relevant nugget is "The German government yesterday said that the root cause of a sharp increase in right-wing attacks against foreigners was the recent surge in the number of asylum-seekers in Germany". For each query, a sample of actual TREC relevant documents was shown to an assessor who extracted relevant nuggets of information in the form of sentences. This set of nuggets was then used to infer the relevance of all unjudged documents containing these nuggets. This process seems impractical since it relies on human assessors to extract nuggets, this is a time consuming task, additionally, the sample of documents to be judged relevant is fixed at 200. Rajput et al. solve these problems by adapting 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 [9]. A distance based approach was suggested by Mollá et al. [19] to expand the number of positive judgments, which means the number of relevant documents for a given query. Based on the cluster hypothesis, which assumes that: "documents in the same cluster behave similarly with respect to relevance to information needs" [16], they computed the distance between the documents and the closest qrel (document judged as relevant by the human assessor). The distance metric used was 1-cosine similarity. The experiments they conducted showed that the distance between a document and a known relevant document can be used as an indicator of relevance. They expanded this work to build "pseudo-qrels" for a given test collection. For each of the weighting models provided by Terrier [15], they selected the top K closest documents retrieved to a known qrel relevant document. To evaluate the quality of the obtained pseudo-qrels, they ranked the 16 different Terrier weighting models using the original set of qrels and then they ranked the same models using the pseudo qrels. There was a positive rank correlation obtained between the two. However, the choice of the known qrels and their number affects the quality of the pseudo qrels. Selecting 20% of the known qrels led to an 80% correlation value for the Kendall tau metric. So, this method still requires human intervention in judging a few documents as relevant before more relevant documents can be obtained.

The work done in this paper is an attempt towards building a relevance judgments lists (which we call new_qrels) with no user intervention, based on automatic information extraction and text matching between key phrases and test collection topics.

Methodology

The steps we follow in our methodology are detailed in this section. As shown by Soboroff [4], when a query is submitted to different Information Retrieval Systems (IRS) and a document is retrieved by all of them, it is most likely that this document is relevant to this query. We call this document as common to the different (IRS).

In our methodology, we submit a query to different weighting models rather than to different (IRS). Example of such models is BM25, PL2 provided by Terrier. This initial query retrieves documents. The common documents among all weighting models are then collected in a set S. Afterwards; an extraction algorithm goes through all the documents in set S to extract key phrases. We select the best key phrases (based on a score described in the subsequent section) and put them in a set called Q. The key phrases in Q are submitted as queries to the BM25 weighting model. We then combine the union of the documents retrieved by the key phrases in Q. These documents are now considered as the newly generated relevance judgment list (qrels) for the initial query; we call them new_qrels.

The advantage of the above methodology is that the new_qrels are formed with minimal human intervention.

Experimental Design

The total number of different weighting models used is twelve. All of them are provided by Terrier. These models are: BM25, DFR_BM25, LGD, In_expC2, In_expB2, IFB2, TFIDF, LemurTF_IDF, PL2, BB2, DLH13 and DLH. Query expansion was carried out for each of the models using 10 documents, adding 40 terms to the query. The query expansion mechanism extracts the most informative terms from the top-returned documents as the expanded query terms. In this expansion process, terms in the top-returned documents are weighted using a particular DFR term weighting model. Currently, Terrier [15] deploys the Bo1 (Bose-Einstein 1), Bo2 (Bose-Einstein 2) and KL (Kullback-Leibler) term weighting models. The DFR term weighting models follow a parameter-free approach in de-fault [16].

The queries submitted to the different weighting models are a combination of the *title* and *description* fields in the original TREC topics [12]. For each query, we select the top k common documents among all weighting models and collect them in a set S. In our experiment, k is set to 10.

The extraction of key phrases from set S is done using the Keyphrase Extraction Algorithm (KEA) [10]; KEA is a plug-in to GATE [11]. The number of key phrases, minimum and maximum number of terms per key phrase can be defined, so for TREC-8, these values were set empirically to 25 key-phrases consisting of 3-5 terms. As for TREC-7, the number of terms was set between 2 and 3.

In this way, each TREC topic will produce a set Q. The key phrases in each set Q are assigned a score by matching them to the corresponding TREC topic; we used the title and description fields to match the key phrases obtained and assign a score to them. We simply counted the number of terms in the key phrase found in the topic title or description. Any

ISSN:2319-7900

key phrase with a score >= 0.4 for TREC-8 and >= 0.33 for TREC-7 was selected as a new query for that topic. It might add more terms to the query or it might show a new aspect of the same topic.

The key phrases in Q are submitted as queries to the BM25 weighting model. We then combine the union of the documents retrieved by each key phrase in Q to form the new_qrels for the TREC topic.

To assess the relevance of each document obtained in the new_qrels, we used the original TREC assessment which was achieved by human assessors. An automatic process to determine the document relevance is a future work. The evaluation of the new_qrels is done using the ranking principle [17][18] which is based on the following: The TREC topics are submitted to a weighting model then two MAP values are computed using the trec_eval package. The first MAP is based on the original qrels and the second MAP that we call new_MAP is based on the new_qrels. This is repeated for twelve Terrier weighting models. Therefore, we get twelve MAP values based on the qrels and another twelve new_MAP values based on the new_qrels. The models are ranked based on the MAP values. Therefore, we obtain two different rankings one based on MAP and the other on new_MAP. A correlation between the two rankings is computed to determine the strength of the relationship between them. The above described process is also repeated for the TREC best short adhoc systems [13, 14].

Experiments and Results

The MAP and new_MAP values obtained for each of the Terrier weighting models for TREC-7 and TREC-8 are listed in tables 1 and 2. For both TREC collections, we can see that the MAP resulting from new_qrels in better than using actual TREC qrels, this is due to the fact that we representing a topic with several aspects or key phrases will retrieve new relevant documents that might not have been retrieved initially. In addition, taking the union of all the queries will lead to more documents than simply using one topic so the probability of finding more relevant documents also increases.

Table 1. TREC-7 MAP and new_MAP values for Terrier retrieval models

	BM25	DFR_BM25	BB2	LGD	PL2	In_expB2
MAP	0.2508	0.2662	0.2637	0.2511	0.2577	0.2645
new_MAP	0.3101	0.3197	0.3156	0.2957	0.3151	0.3143
	In_expC2	IFB2	DLH	DLH13	TF_IDF	Lemur TF_IDF
MAP	0.2559	0.2629	0.2678	0.2691	0.2515	0.252
new_MAP	0.3106	0.3146	0.3229	0.3183	0.3113	0.3019

Table 2. TREC-8 MAP and new_MAP values for Terrier retrieval models

	BM25	DFR_BM25	BB2	LGD	PL2	In_expB2
MAP	0.2901	0.2918	0.2894	0.294	0.2721	0.29
new_MAP	0.3482	0.3605	0.359	0.3616	0.3358	0.3596
	In_expC2	IFB2	DLH	DLH13	TF_IDF	Lemur TF_IDF
MAP	0.2805	0.2903	0.2898	0.2997	0.2876	0.2708
new_MAP	0.3474	0.3146	0.3556	0.3765	0.3576	0.3306

We compare both MAP and new_MAP for the best automatic short adhoc systems reported in each of the TREC-7 and TREC-8 results. Similarly to the results obtained for the Terrier models, the new_MAP values seem to outperform the MAP values for TREC automatic systems as shown in tables 3 and 4.

Table 3. TREC-7 MAP and new_MAP values for TREC best automatic systems

	att98atdc	bbn1	Cor7A3rrf	INQ502
MAP	0.2961	0.2797	0.2674	0.2815
new_MAP	0.3126	0.296	0.2711	0.2856
	mds98td	ok7ax	pirc8Aa2	tno7exp1
MAP	0.2809	0.3033	0.2723	0.2785
new_MAP	0.3031	0.3058	0.2733	0.274

Table 4. TREC-8 MAP and new_MAP values for TREC best automatic systems

	att99atde	Flab8atd2	fub99td	ibms99a
MAP	0.3165	0.293	0.3064	0.3005
new_MAP	0.3257	0.3007	0.3232	0.3145
	MITSLStd	ok8amxc	pir9Aa1	tno8d3
MAP	0.2979	0.3169	0.2624	0.2921
new_MAP	0.3063	0.3318	0.241	0.3102

Following Efron, our main way to evaluate the results we obtained was to evaluate the correlation between the system rankings obtained for the original TREC qrels and the new_qrels obtained by our method, both for a range of Terrier [9] retrieval models and a set of automatic short query ad hoc TREC systems. For this purpose, we computed the Spearman coefficient.

Other than comparing the correlation with benchmark systems, we also compare our results with those obtained by Efron using multiple aspects to build pseudo-qrels. In his method, Efron created several aspects for each TREC topic and then using a single seed system Okapi, he ran each aspect and then created the pseudo-qrels by taking the union of the top 100 documents obtained from the aspects of each topic. He called the measure obtained aMAP (aspects MAP). In Table 5, the second column shows the rank correlation between MAP and new_MAP while the third column shows the rank correlation between MAP and aMAP.

ISSN:2319-7900

Table 5. Rank correlations using new_MAP and aMAP

Data	Spearman for TREC automatic systems using new_MAP	Spearman for TREC systems using aMAP (query aspects)
TREC-7	0.904	0.974
TREC-8	0.928	0.92

As for the Terrier models the Spearman coefficient obtained using new_MAP was 0.909 for TREC-8 and 0.881 for TREC-7. The correlation between MAP and aMAP was slightly higher than that obtained between MAP and new_MAP for TREC-7, but there was very little difference for TREC-8. The main difference between Efron's technique and ours is that no human effort is required in our method to create the aspects of each topic. Also, in some cases, the aspects used by Efron used some additional information extracted from Wikipedia or a dictionary whereas the key phrases are matched exactly with the TREC topic in our method.

Conclusions

In this paper, we provide a new technique for automatically generating a set of qrels using minimal human effort. We evaluate the results obtained by comparing system rakings of different Terrier models and the benchmark TREC systems. With a correlation >= 0.9, we can say that the new_qrels obtained are reliable and can be used for evaluating retrieval systems.

References

[1] Cleverdon C.W. and Keen E. M. (1966). "Factors Determining the Performance of Indexing Systems", N. Cranfield, UK: ASLIB Cranfield Research Project.

[2] Sparck Jones K. and Bates R.G., Report on a design study for the `ideal' information retrieval test collection, Computer Laboratory, University of Cambridge, 1977 (BL R&D Report 5428).

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

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

[5] 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.

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

[7] 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 (2004) 33-40

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

[9] 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

[10] Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G. "KEA: Practical automatic keyphrase extraction." Working Paper 00/5, Department of Computer Science, The University of Waikato, 2000.

[11] H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.

[12] Vorhees, E.M: Evaluation by highly relevant documents. In SIGIR'01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM, 2001. 74-82

[13] Harmon, D.K., Vorhees, E.M.: Overview of the Seventh Text Retrieval Conference (TREC-7). DIANE Publishing Company (1996)

[14] Harmon, D.K., Vorhees, E.M.: Overview of the Eighth Text Retrieval Conference (TREC-8). DIANE Publishing Company (1996)

[15] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald and Douglas Johnson. Terrier Information Retrieval Platform. In Proceedings of the 27th European Conference on Information Retrieval (ECIR 05).

[16] Manning C.,Raghavan P., and Schütze. An H. Introduction to Information Retrieval, Cambridge University Press, Cambridge, England 2009

[17] Robertson, S.E. (1977). The probability ranking principle, in: IR Journal of Documentation, 33(4), pp. 294-304

[18] van Rijsbergen, C.J. Information Retrieval, Butterworth-Heinemann, Newton, MA, 1979.

[19] Mollá D, Martinez D. and Amini I. Towards Information retrieval evaluation with reduced and only positive judgments, ADCS'13, ACM 978-1-4503-2524-0, 2013