Exploration of Term Relevance Sets

Bjørn Harald Olsen
bjornhol@idi.ntnu.no

Jeanine Lilleng
jeanine.lilleng@idi.ntnu.no

Department of Computer and Information Science
Norwegian University of Science and Technology

Abstract

Evaluation of information retrieval (IR) systems can not always be based on the established pooling techniques and the resulting test sets. This especially affects domain specific IR systems and IR systems tailored for languages not supported by the test sets.

IR system evaluation based on Term Relevance Sets (Trels) represents an alternative for these IR systems. The technique is still novel and little documentation exists. This paper presents the technique and our first investigations into Trels.

Introduction

Information retrieval (IR) deals with making information easier available. Well known applications of information retrieval are search engines like Google [2] and Alltheweb [3]. Other examples of information retrieval are document management systems in companies. Information retrieval systems are evaluated based on how well they satisfy information needs. The evaluation can be performed in different ways. This paper discusses an alternative approach to the main stream methods commonly used for information retrieval system evaluation.

Initially, IR system evaluation was based on manual relevance judgment of every document contained in the collection. Relevance is defined as the documents ability to cover the specified information need. This approach was suggested and investigated during the Cranfield tests [4]. As the collection size grew and the number of IR systems increased, the amount of manual work required to continue this approach became prohibitive.

The pooling technique as described by Spark-Jones and van Rijsbergen [5] was used to make larger test collections possible. Large IR system evaluation efforts like TREC [6, 7], NTCIR [8] and CLEF [9] are based on this technique. Pooling reduces the necessary effort put into relevance judgment of the documents in the test collection, but the trade off is an increased need for IR system runs to construct the pools [10].

Term Relevance Sets as basis for evaluation of IR performance, is a novel approach recently presented by Amitay et al [1]. The main idea is to use term occurrences to perform an automatic relevance judgment of documents. This approach for evaluation is especially attractive for developers of IR systems designed for use with languages not involved in TREC, CLEF and NTCIR or in domains not covered by these initiatives.

Even though the main principles of Terms Relevance Sets are presented in Amitay's paper, the main focus is on showing that Trels-based evaluation gives comparable results to using the test sets of TREC. It does not discuss or test the necessary size of the Term Relevance Sets, or which value of a tScore, measure used to indicate relevance that indicates a relevant document. This paper presents an initial investigation into both tScores and the size of Term Relevance Sets. This paper also gives an introduction to IR system evaluation using Term Relevance Sets and reports on our initial investigation into this approach.

This paper was presented at the NIK 2006 conference. For more information see http://www.nik.no/

Related work

TREC and Trels are examples of IR system evaluation methods, where the user experience is abstracted away. The relevance of the retrieved documents are automatically calculated by the respective evaluation method. The work done by Amitay et al [10] is the basis of the work presented in this paper. The authors are not aware of other literature concerning Trels. The main research effort in IR system evaluation today seems not to be put into system evaluation methods, but into the more difficult challenge of user-based evaluation methods.

Challenges in standard IR evaluation systems

Even though the methods used today in large evaluation efforts like TREC are less labor-intensive than the early Cranfield tests, thousands of documents still must be given a manual relevance judgment during the preparation of a new test collection. The requirement for a large number of test runs makes this method for IR system evaluation unrealistic for some groups. IR systems can be tailored for specific domains or languages that are to narrowly populated by IR systems to justify this kind of evaluation effort.

Some issues also exist concerning the assumptions that the pooling technique is based on. The first being the effect that difference in human judgment has on the test collections. Normally only one “expert” evaluates whether a document is relevant or not. Experiments done by Voorhees [9], during TREC 4, show a disagreement in experts’ judgment of document relevance for the same document. However, the results indicate that the effect of this disagreement on final IR system ranking is reduced as the number of topics increase.

Another challenge is the relevant documents not found by any IR system. These are considered irrelevant, and are never evaluated for relevance. Voorhees [9] argues that this will not affect the comparability of the TREC runs. She points out that other evaluations based on few test runs and few IR systems than TREC will be more vulnerable to this problem.

Most IR systems are based on the same basic principles. However, IR systems based on other principles, evaluated towards existing test runs (collections), might make comparison difficult because relevant documents only found by the new system are considered irrelevant due to the pooling technique.

What makes TRELS interesting

For evaluation of general IR systems designed for supported languages (TREC [6]; English, CLEF [9] & NTCIR [8]; other languages), the available test sets will give a good indication of retrieval performance. For other languages these test sets will not be usable. However, Trels-based evaluation can still be used. The 0.993 correlation found for the 128 participants in TREC-8, when comparing the Trels-based evaluation and the TREC evaluation of the systems [10], indicates that Trels can be a good alternative.

The fact that created Trels can be reused on other collections and scale well, can be exploited to perform IR system performance evaluations on both large and dynamic collections, e.g. the Internet.

Trels

The idea behind Trels-based evaluation is to use the occurrences and absences of terms in a specific set to perform the relevance judgment for a specific document. This set of terms is called the *Term Relevance Sets*.

The Term Relevance Sets (Trels) consists of two different term sets: *On Topic Terms* and *Off Topic Terms* which are related to *a* specific *query*.

- On Topic Terms: Terms related to the query that are **likely** to appear in relevant documents.
- Off Topic Terms: Terms related to the query, but are **unlikely** to occur within relevant pages.

Basically, occurrences of On Topic Terms and absence of Off Topic Terms indicates relevance. On the other hand; occurrences of Off Topic Terms and absence of On Topic Terms indicates that the document is irrelevant to a specific query. A sample Trels is given below:

```
query: "recycle,
automobile tires"

 onTopic: "rubberized
asphalt", "door mats",
playground

 offTopic: "traction, air-
pressure, paper, plastic,
glass
```

An example Trels [1]

As known, old tires are recycled into door mats or rubberized asphalt and they are often used at playgrounds. To identify good On Topic Terms we either use prior knowledge of terminology in the relevant documents, or we could use a third-party retrieval tool to browse through documents returned based on the query. A third-party retrieval system can also be used to find Off Topic Terms. Often occurring terms in irrelevant documents returned by the IR system can be used. Further, neither of the Topic Terms should be individual query terms and linguistic derivates of the query terms.

tScore calculations

The occurrence of On Topic and Off Topic Terms are counted to give the tScore of a document. The tScore is a measure of relevance. The higher the tScore, the more relevant is a document considered to be. These individual document tScores are then combined to give the tScore of a specific returned set of documents.

The tScore for retrieval results for query q, consisting of n document d, is done in two steps; first tScore(d_i ,q) for each retrieved document d_i is calculated. Then these results are combined to a tScore(Dq), reflecting all the documents. The figure below illustrates both the document level scoring and aggregated retrieval score.

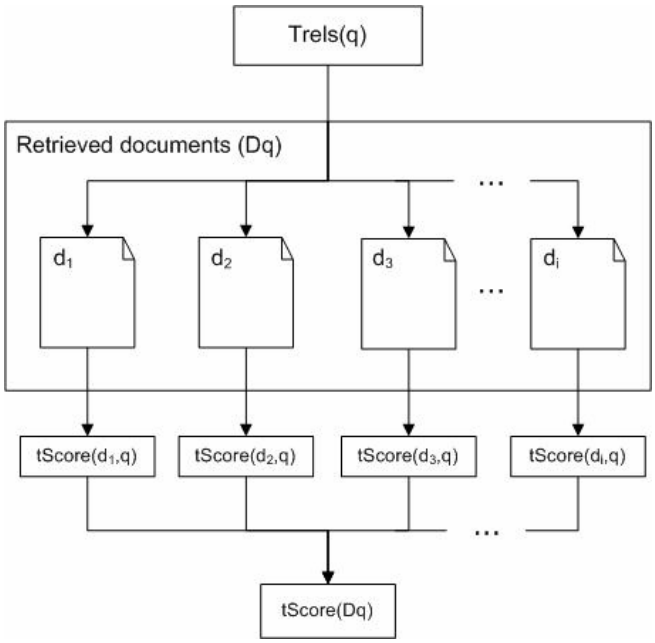

Figure 1: Document level scoring *tScore(d_i, q)* and collection scoring *tScore(D_q)*

There are two different schemes for calculating tScore(d,q) of a document d; equation 1 or equation 2. The first scheme simply counts the number of occurrences of On Topic and Off Topic Terms. The Off Topic Terms can be given a higher or lower relevance compared to the On Topic Terms by giving β a value above or below one.

$$tScore_{basic}(d,q) = |\, t \in onTopic \cap d \,| - \beta \times |\, t \in offTopic \cap d \,| \quad \text{(eq. 1)}$$

The second method for calculation of tScore(d,q) is called the *Similarity scheme* and is based on the principles of the vector model. The measure is found by comparing the document vector to both a vector based on the On Topic Terms and a vector based on the Off Topic Terms.

$$tScore_{sim}(d,q) = \cos(onTopic, d) - \beta \cos(offTopic, d) \quad \text{(eq. 2)}$$

Experiments gave a correlation of 0.991 between the two schemes, and the choice of scheme depends on facilities supported by the IR environment. The constant β is set according to how heavily the Off Topic Terms should be weighted. Setting β = -1 makes the Off Topic Terms count as On Topic Terms. It seems reasonable to set β = 1, which is done for the experiments in this paper, because it gives the highest correlation to retrieval performance found using TREC-8 test collection.

To calculate the score for a result set D_q for query q either equation 3 or equation 4 can be used to give the tScore(Dq).

$$tScore(D_q) = \frac{\sum_{i=1}^{n} \frac{1}{i} tScore(d_i, q)}{\sum_{i=1}^{n} \frac{1}{i}} \quad \text{(eq.3)}$$

$$tScore@k(D_q) = \frac{1}{k} \sum_{i=1}^{k} tScore(d_i, q) \quad \text{(eq. 4)}$$

Equation 3 and 4 differs in their emphasis on the documents ranking. In equation 3, higher ranked documents contribute more to the tScore than lower ranked documents Equation 4 calculates the mean values of all the tScore(d_i,q), giving higher and lower ranked documents the same weight.

No normalization of tScore based on document length was suggested by Amitay et al. We find document length normalization of tScore important to avoid that longer documents "automatically" are considered more relevant than shorter documents. Without length normalization, a 100 words document consisting of 10 On Topic Terms and 0 Off Topic Terms would get a tScore of 10 using the basic scheme. If the body of the document were duplicated, the number of On Topic Terms would be duplicated as well, giving a tScore of 20. Hence, we suggest extending both equation 1 and 2 with normalization giving equation 1n and 2n. We have used the normalized version of equation 1, i.e. 1n in our experiments.

$$tScore_{basic}(d,q) = \frac{1}{\#terms} \times \left(\left| t \in onTopic \cap d \right| - \beta \times \left| t \in offTopic \cap d \right| \right) \quad \text{(eq. 1n)}$$

$$tScore_{sim}(d,q) = \frac{1}{\#terms} \times \left(\cos(onTopic, d) - \beta \cos(offTopic, d) \right) \quad \text{(eq. 2n)}$$

We expect the number of On Topic and Off Topic Terms to significantly affect the accuracy of evaluations using Trels, but we have found no discussions on this in the literature. An average number of 32,5 On Topic and 8,5 Off Topic Terms were used in [10], but no discussion on how these numbers where provided. Having normalized the document tScores, we also expect there to be a threshold value that can be used to determine if a document is relevant to the specific query. The experiments described below are designed to be initial investigations into these issues.

Experimental set up

Issues we want to explore:

- How many terms are needed in on/off topic sets.
- Exploring expected tScore threshold for relevant and irrelevant documents.

For these initial investigations two different Trels with corresponding document collections were constructed. For both collections 20 On Topic Terms and 20 Off Topic Terms were created. The terms considered most important for relevance judgment were placed first in the sets.

For each Trels a collection of relevant and irrelevant documents representing a fictive document retrievals were constructed. The collections were manually created to minimize the influence any on IR system would have on the results. The documents` relevance was manually judged.

Test case 1

Query: "hvordan oppdra barn" (in English: how to raise children)

On Topic Terms barneoppdragelse, mamma, pappa, autoritet, kjærlighet, grenser, straff, respekt, familie, vilje, sinne, gi etter, konflikt, misunnelse, forklare, irettesette, streng, snill, utvikling, ungdom ***Off Topic Terms*** dyr, hund, hest, valp, føll, sykdom, vannkopper, underholdning, bøker, svangerskap, adhd, rode hunder, meslinger, feber, bleier, kosthold, gravid, skillsmisse, barneklær, separasjon

Table 1: Content of Trels test case 1

Collection

All the documents found in this collection are probable returned results when using the specified query in a search engine. The documents judged irrelevant are mostly on "raising" dogs. Some irrelevant documents also discussed children in general. The irrelevant documents are considered to be somewhat related to the relevant documents, in this collection.

Query	hvordan oppdra barn
Relevant documents	10
Irrelevant documents	10
Average document size	5.9kB
Relevant document subjects	Tutorials on how to raise kids
Test purpose	Separate a subject inside a domain

Table 2: Test case 1

Test case 2

Query: "springer" (in English: bishop (in chess))

On Topic Terms brettspill, spill, oppstilling, angrep, forsvar, svart, hvit, løper, konge, droning, tårn, bonde, hest, sjakkmatt, matt, rokkering, taktikk, motstander, brett, koordinat ***OffTopic Terms*** trene, jogge, kondisjon, arkiv, hund, sykle, verlag, spaniel, engelsk, kennel, oppdrett, trening, mat, axel, løype, skog, rase, mosjon, sykle, fuglehund.

Table 3: Content of Trels test case 2

Collection

The irrelevant documents were mostly about dog breeds (springer spaniel). There were also collected irrelevant documents that were about exercise ("springer" has the second meaning of "running" in Norwegian). Relevant documents were picked from Web pages containing chess rules, match reports and tutorials. The irrelevant documents in this collection were chosen to be from a different domain (topic) than the relevant documents.

Query	springer
Relevant documents	10
Irrelevant documents	10
Average document size	8.5kB
Document subjects	Chess rules, tips and tricks, match logs

Test purpose	Differentiate documents from different domains

Table 4: Test case 2

The runs

Run 1

This run was expected to give information of how the number of Terms used effects the tScore of the system. We also wanted to investigate whether equation 3 and 4 gives different aggregated tScores.

Only the relevant documents, from both collections, were used. The weighted aggregated and the average tScore were calculated as the numbers of included On Topic and Off Topic terms were varied. We chose to keep the number of On Topics and Off Topic Terms the same during the whole of our experiment.

Run 2

As we decrease the number of On Topic and Off Topic Terms we expect the order of the Terms to matter. This run was designed to investigate how the order of the Terms affects the tScores.

The tScores were calculated for the relevant documents in both test sets. For each test case three different tScores were calculated: One with the terms ordered with expected best Terms first. The last graph represents an average over 200 calculations, where the order of the terms is random.

Run 3

This run was initially designed to find the threshold value for relevant documents, but it also gave relevant information about the necessary size of On Topic and Off Topic Term relevance sets.

For both test sets three different aggregated tScores were calculated: One based only on the relevant documents in the collection. One based only on the irrelevant documents. The final aggregated tScore were based on a document set where half the document were relevant, and half were irrelevant. In this experiment the term relevance sets were ordered with the expected best terms first.

Results

Run 1

The results of this first run can be found in Figure 2 and Figure 3. Both equation 3 and 4 show monotone increasing tScore as the number of included On Topic and Off Topic Terms increased. This means that the confidence of the documents being relevant increases with use of several terms. Since we know that the documents are relevant, this is an expected outcome.

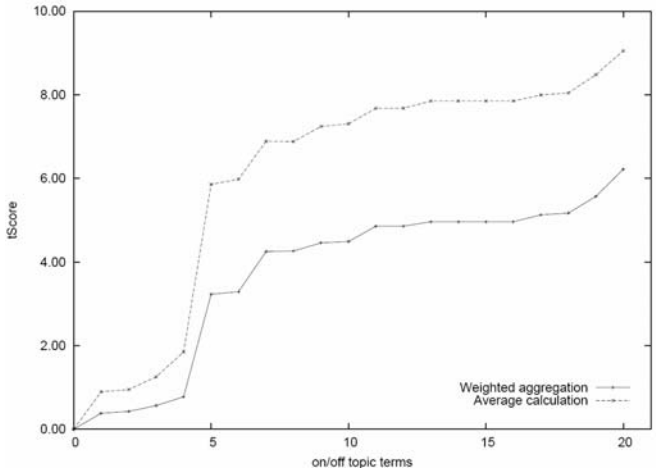

Figure 2: Run 1, test case 1

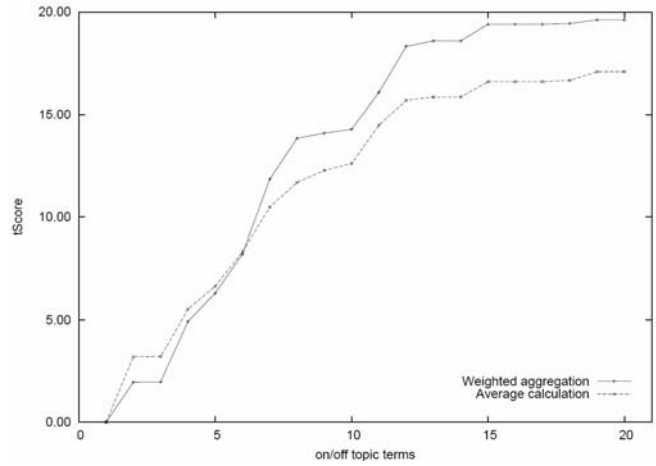

Figure 3:Run 1, test case 2

If we take a look at the difference between equation 3 (weighted aggregation) and 4 (average calculation) for both test cases, we see that in test case 1 the average measure gives the highest score, but for test case 2 the opposite (when using more than 6 terms) can be observed. “Weighted aggregation” weights the top-ranked documents heavier than lower ranked documents. This gives an idea of how well the documents are ranked. Figure 2 and 3 shows that the document ranking used is less good in test case 1, than in test case 2.

The differences between the two ways to calculate aggregated tScore seems to give relative comparable results. Different results might have been found if also irrelevant documents had been included in the experiments, but this was not done. This might reduce the validity of the results.

For the two remaining runs only the aggregated tScore based on average document tScores (equation 4) will be used. Equation 4 calculates the mean score of the top k ranked documents. This is expected to make it easier to compare results with measurements based on recall and precision since they also are based on the k highest ranked documents. Our k value equals the number of documents in the collections.

Run 2

If an "expert" uses domain knowledge to create Trels, she would probably write down the first terms that come to her mind first. This may result in a set of terms with

decreasing degree of importance. Previously, it was found that 10 terms is enough to determine relevancy of a document collection. Yet, it is tried to change the order of terms added, so that the most important terms are applied first. Additionally, the inverse order was plotted, so the terms come in increasing degree of importance. This was expected to tell if the optimal term choice (among the ones originally used) reduces the number of terms needed.

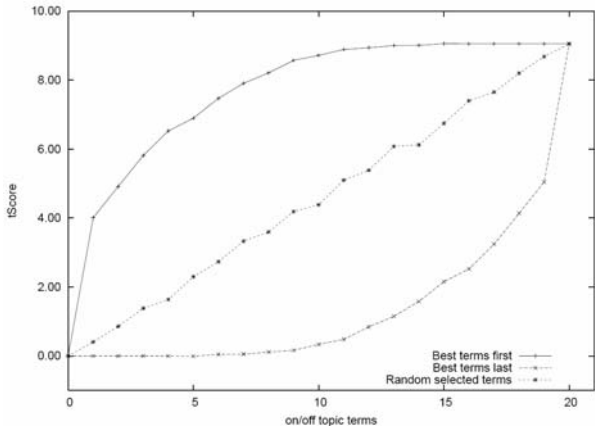

Figure 4: Run 2, test case 1

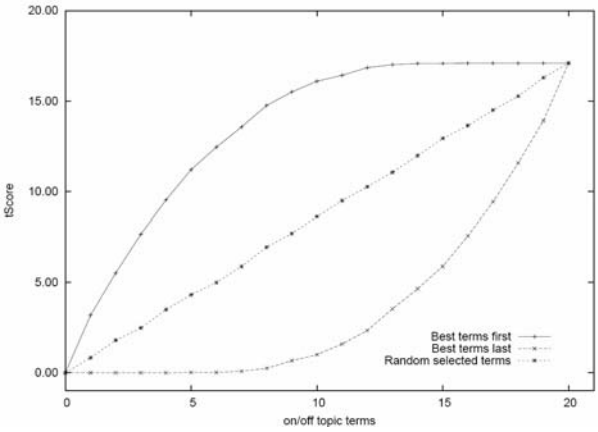

Figure 5: Run 2, test case 2

Figure 4 and 5 shows that we reach the maximum level of score after roughly 10 terms if the most significant terms are chosen first. The original term set, where "own knowledge" were used to pick terms, also gave the same conclusion. This tells us that using an expert's knowledge worked as good as forcing the best terms to come first. In the second plot in figure 4 and 5 where the best terms are picked last, show that after 10 terms we can hardly tell if these documents are relevant or not. The plot labeled "random selected terms" is the result of picking x terms randomly N=200 times and calculating the average tScore. This basically shows that, if the terms are picked without any directives, all 20 terms are needed to reach the same level as choosing terms either with "own knowledge" or forcing the best to come first.

Run 3

Our initial theory that a tScore above zero indicates a relevant document holds for both test cases when there are more than 5 On Topic and Off Topic Terms. The graph representing the aggregated tScore for the result with both relevant and irrelevant

documents is quite close to the zero axes for both runs. This makes sense since we use equation 4 that gives all documents the same weighting.

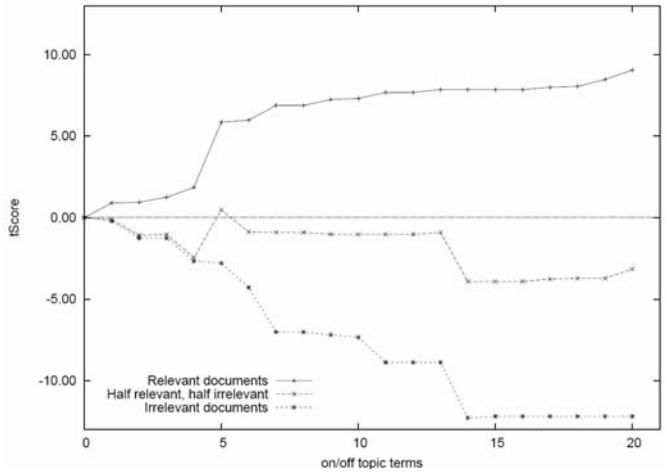

Figure 6: Run 3, test case 1

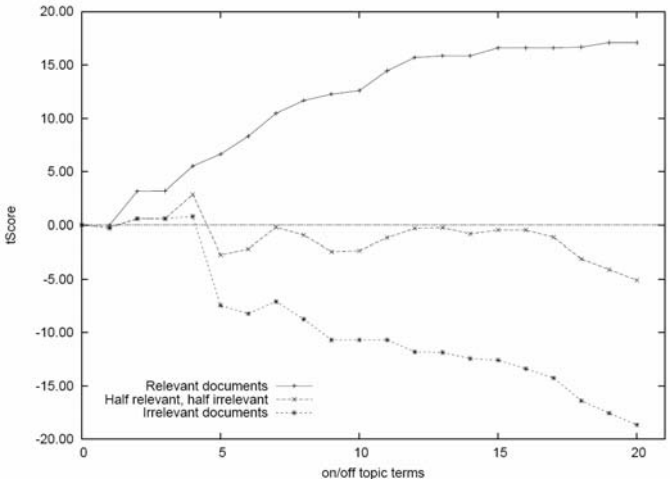

Figure 7: Run 3, test case 2

Both test cases in run 3 shows that there are little difference between the tScores of the relevant and the irrelevant documents when the number of On Topic and Off Topic Terms are less than four terms. The strange deviations on both graphs around 5 On/Off Topics Terms might be caused by bad choice of Off Topic Terms. This could have been evaluated by doing the second run also on Off Topic Terms.

Still, as the number of On Topic and Off Topic Terms increase a tScore of 0 seems to be neutral, indicating the same amount of relevant and irrelevant documents. We expect that this threshold can be used on individual documents to indicate relevance / not relevance for a query.

Obviously, there is a trade-off between the number of On Topic and Off Topic Terms used, the certainty of the tScore, and the necessary manual work to create the on/off topic terms. Using four on/off topic terms seems to result in "random" relevance judging of documents. Using 5-6 on/off topic terms for calculating tScore, increases the reliability of the prediction significantly. After ten on/off topic terms the certainty of the prediction as the number of on/off topic terms increase seems be leveling. Since it becomes more difficult to find good on/off topic terms as the number increases, and little effect was discovered, 10 On Topic and 10 Off Topic Terms seems to be a good cut of point in our preliminary exploration into Trels.

Discussion

Most of the results in the runs were as expected, with two exceptions: The suggested amount of necessary On Topic and Off Topic Terms, and the way the graphs in run 2 seems to indicate something about the quality of the Trels.

In the paper that presents this approach an average of 32.5 On Topic Terms and 8.5 Off Topic Terms were used. Our results indicate that this number of terms might be higher than necessary. One of the main advantages with Trels is the way they reduce necessary manual work to evaluate IR systems that cannot use pooling based on test collections. If the number of terms can be reduced even further to a total of 20 (10 On Topic Terms and 10 Off Topic Terms), this might be more than 50% reduction in manual work necessary to prepare Trels.

Individual judgment on how many On Topic and Off Topic Terms, for each Trels, increases the amount of personal judgment that goes into creation of Trels. This might make the Trels and the evaluation of relevance based on them less objective.

One of the nice features with Trels is the possibility for reuse. If we were able to find objective and automatic techniques that can evaluate the quality of the On Topic and Off Topic Terms, this would hopefully make them less subjective. Run 2 shows the worst order, best order and random order of terms. As you can see from figure 4 and 5, a reduction in tScore can be observed when a term that does not help to indicate relevance is added.

If you have a small test collection with relevant and irrelevant documents, we believe that this kind of graph can be used to improve the quality of Trels before they are used for IR system evaluation.

Conclusion & Further work

We believe that Trels based evaluation is a promising technique for evaluation of some IR systems. This especially applies for IR systems tailored for languages not covered by large pooling efforts, like Norwegian, and IR systems for use in narrow domains where no test collections exists.

Our initial investigation indicates that 10 On Topic and Off Topic Terms might be sufficient to predict relevance / non relevance of the documents. If this can be validated in a larger study, creation of Trels is even less labor intensive than indicated by Amitay et al.

No cut of value for tScore for relevance were suggested in the initial paper on Trels. The cut of value is necessary to used Trels for actual IR systems evaluation. In our initial exploration into Trels a threshold value of zero seems to give good automatic relevance judgments.

All our results are based on a very small amount of tests, and can therefore only be regarded as indications on how Trels will behave in real IR system evaluations. To be able to validate our findings properly, we are now moving on to repeat the experiments on a much larger scale.

We also believe that run 2 shows some promising results that might eventually result in a quality assurance method for Trels. If the graphs are created based on very small well established test set, we expect that terms not contributing to the tScore and terms that are contra productive can be removed from the Trels, hence improving their quality.

References

[1] Amitay, E., et al, **Scaling IR-system evaluation using term relevance sets,** Proceedings of the 27th annual international conference on Research and development in information retrieval, Sheffield, United Kingdom, 2004, 10-17

[2] **Google search engine** www.google.com [19.09.2006]

[3] **Alltheweb** www.alltheweb.no [19.09.2006]

[4] Cleverdon, C. W., **The significance of the Cranfield tests on index languages,** Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, Chicago, Illinois, United States, 1991

[5] Sparck Jones, K. and van Rijsbergen, C., J., **Report on the need for and the provision of an 'ideal' information retrieval test collection,** Cambridge, 1975,

[6] **TREC Text Retrieval Conference** http://trec.nist.gov/ [02.03.2005]

[7] Voorhees, E. and Harman, D., **TREC: experiment and evaluation in information retrieval / edited by Ellen M. Voorhees and Donna K. Harman**, Cambridge, Mass.: MIT Press, 0-262-22073-3, 2005

[8] **NTCIR** http://research.nii.ac.jp/ntcir/ [18.06.2006]

[9] **CLEF Cross Language Evaluation Forum** http://clef.isti.cnr.it/ [02.03.2005]

[10] Voorhees, E. M., **The Philosophy of Information Retrieval Evaluation**, 2002