TREC2004 Robust Track Experiments using PIRCS

K.L. Kwok, L. Grunfeld, H.L. Sun and P. Deng

Computer Science Department, Queens College, CUNY
Flushing, NY 11367

1 Introduction

There were two sub-tasks in the TREC2004 Robust track: given a set of topics, a) improve the effectiveness of the lowest performing 25%, and b) predict their ranking according to their average precision.

For task a), we followed the strategy introduced by us last year to improve ad-hoc retrieval by employing the web as an external thesaurus to supplement a given topic description. A new method of probing the web based on a given topic statement called 'window rotation' was tested.

For task b) we employed ε-SVR (epsilon support vector regression) to predict performance of test topics based on training with some simple features such as document frequencies, query term frequencies. This allows performance prediction without retrieval. Features were also added from a retrieval list with the hope that they may predict later stage or web-assisted retrieval better. 200 old topics were used for training to predict the ranking of 49 new topics, as well as the whole set of 249.

Runs were done that made use of title only, description only section of a topic, and title-description-combination retrieval lists. Ten submissions including runs that were based on initial retrieval only, retrievals with pseudo-relevance feedback, and with web-assistance. Evaluation shows that we have achieved very good performance for most of our runs.

2 Robust Track – Improving Low Performing Topics

2.1 Background

We introduced a new strategy of improving ad-hoc retrieval based on web-assistance in the Robust Track of TREC2003. In initial retrieval, some queries have low average precision performance (weak or hard queries) while others return good values (strong or easy queries). The objective of this track is to automatically improve the effectiveness of weak topics, and others in general. Strong topics can generally be further improved with pseudo-relevance feedback (PRF), but this does not work for weak topics because for them, an initial retrieval would not bring in much useful material for feedback use. One may try to enrich weak topic wordings via a thesaurus to improve term variety, and thereby enhancing initial retrieval results. However choosing an available and appropriate thesaurus of the right domain without prior knowledge of a topic is quite a challenge. We demonstrated in TREC2003 that employing the WWW as an all-domain word-association resource with appropriate filtering can be successful for this Robust Track objective.

Additional to normal ad-hoc retrieval on the target collection with the original TREC topics, our method of employing web-assistance consists of four steps, and these are illustrated in Fig.1:

1) for each TREC topic, define associated web queries for a specific search engine;

2) use the web queries with the search engine to probe the web for relevant or related pages or snippets;
3) from the returned web pages or snippets, define alternate queries (based on proportional word frequencies) for retrieval from the target collection;
4) combine retrieval lists from the original TREC query and the alternate queries to form the final retrieval result.

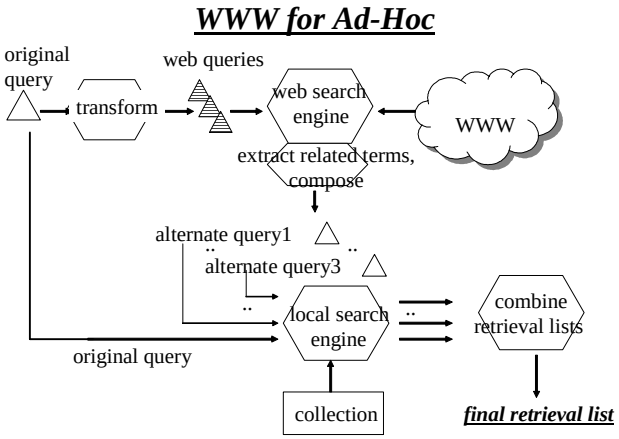

Fig.1: Web-Assisted Ad-Hoc Retrieval

Reasoning backwards, one sees that this approach works if the alternate queries possess synonyms or lexically different content terms but are semantically related to the original. This implies that the returned web pages/snippets should be answers or topically close to what the TREC topic wants. To achieve this, the web queries defined in Step 1 are of paramount importance, and would determine the success or failure of this strategy.

The web queries naturally are tied to which search engine one wants to employ. We have focused on using the Google search engine because it is generally effective, offers a convenient API (although we wrote our own interface), and searches an immense collection (over 4G web pages according to its homepage). Google accepts input queries in Boolean AND form: a AND b AND c .. AND d (where a,b, ..d are un-stemmed single keywords or phrases), although its final retrieval is further dependent on page-link analysis. Even though Google's search collection is huge, the Boolean AND operator reduces the output answers rapidly to null when the number of query terms increases (7 or more for example depending on the terms). If a TREC topic is short, like those composed from the 'title' section with 2 to 5 words, one could just use all the content terms as a web query. If one considers longer queries like those obtained from the 'description' section of a TREC topic which is mostly a sentence long, one has to employ some filtering methods to choose salient terms from the 'description' for web retrieval. Selecting salient terms from a sentence or a piece of text is not an easy task and is prone to error. Attempts for salient term selection include (Dorr, Zajic, Schwartz 2003) for headline generation for example. In TREC2003, we employed word/sentence syntax for this purpose. In addition, data redundancy and fusion was used to improve retrieval effectiveness.

MINIPAR is a broad-coverage parser available from (Lin 1994) that, among other data, tags each input word of a sentence according to its POS, as well as indicating the asymmetric binary relationship between word pairs that play the role of governor and dependent. It also recognizes phrases which we call (MINIPAR phrases). Last year, as a general approach we tried selecting

salient words based on phrases, and nouns, verbs, adjectives in this order until we obtained a fixed number of words (5 or 6), or in addition choose semantic categories of person, country, organization first, or simply just the first 6 content words of a description.

Because of brittleness for Boolean retrieval and the uncontrolled nature of web content, we employ multiple (3) web queries so as to have higher chance that at least some of them would have sufficient salient terms. These queries return web pages/snippets that would be used to define alternate queries for target collection retrieval. We also rely on data fusion: past experience has shown that when combining retrieval lists that are reasonably different from each other, the combined final retrieval tends to get a boost in effectiveness above the component lists.

This same strategy was followed this year, except that we also tested a method of generating web queries from TREC topics without the need for salient term selection. This is the window rotation method to be discussed in Section 2.4.

2.2 Nomenclature for Web Queries, Alternate Queries and Their Retrieval Lists

Different types of web queries can be generated from TREC2004 topics based on sentence syntax that is the result of analysis by MINIPAR, or other analysis. We will employ a systematic nomenclature w-x_y-z to represent them: the first two symbols w-x pertain to properties of a TREC topic, and y-z describes web retrieval properties. Each symbol takes values as follows:

w = {t, d, ..} denotes the source of a query such as t (from title section of a TREC topic), d (description), etc.;

x = {n, v, a, p, w, ..} denotes the word syntactic categories or other properties of the source w that are employed to help define web queries. These include phrases (p), nouns (n), verbs (v), adjectives, content words (w), etc.

y = {s, r, ..} denotes the mode of web retrieval such as single (s), window rotation (r), etc.

z = {s, f, ..} denotes the granularity of retrieved items that can be snippets (s), full page (f), mixtures (sf), etc.;

When no ambiguity arises, the notation will be used to denote the web queries, the alternate queries derived from web retrieval, and the resultant retrieval lists by these alternate queries from the target collection. As can be seen, numerous types of web queries (and their resulting alternate queries and retrieval lists) can be obtained based on the various parameter settings. The runs we have used in TREC2004 are described in the following sections.

2.3 Queries from Title Section of TREC Topic

For retrieval using the title section, the original TREC queries (and their retrieval results) are denoted as t-init and t-prf: for initial and 2nd stage retrieval based on PRF. The initial queries average to about 3.1 terms (with some 2-word phrases), and none longer than 5. Four Google queries are defined based on various title word and web retrieval properties. These are listed below:

t-init -- initial query from the title section of a topic processed with stop-word removal, Porter stemming, and some 2-word phrases.

t-prf(10,90) -- PRF query expansion with 10 top documents and 90 top terms from t-init retrieval.

t-w_f-s(40,60) – content words from the title to retrieve 40 full web pages that define alternate queries with 60-term output. Page sizes are restricted to <30K. Special file types such

as pdf, doc, etc are eliminated (similar to 'qts' of Lazslo et.al. 2003).

t-w_sf-s(40,90) – content words retrieving 40 full pages and their snippets, maximum 90-term alternate queries output. This differs from the previous mainly in adding snippets to full-page web output and defines a different size for the alternate queries.

t-w_s-s(100,60) – content words retrieving 100 snippets; maximum 60-term alternate query output.

t-pw_sf-s(40,90) – MINIPAR-identified 2-word phrases (and used as phrases for web retrieval) plus other content words retrieving 40 web pages with their snippets; maximum 90-term alternate queries.

The above web queries are chosen to provide diverse alternate queries, as well as based on their training results. TREC2003 Robust Track topics are used for training, and results are shown in Table 1 in three sets: 100 topics (All100), New50 (new during 2003) and Hard50, where All100 is the union of the other two sets. The measures used for evaluation include: MAP (mean average precision), P10 (mean precision at 10 documents retrieved), #0P10 (number of topics without relevant documents at 10 documents retrieved) or its percentage #0P10%, and area measure defined in (Voorhees 2004) which can be viewed as a weighted sum of the lowest 25% MAP values, with weights favoring the lower ones.

It can be seen from Table 1 that for the weak queries (Hard50), initial results of t-init is much better than t-prf (2^{nd} stage retrieval) in the 'area' and '#0P10%' measures (although MAP is worse), while the reverse is true for New50. New50 are stronger (easier) queries because results are better than Hard50 for all queries and for all evaluation measures. This suggests that PRF is detrimental to weak queries based on the 'robust' evaluation measures. In all cases, PRF is effective for MAP values, which may be viewed as a measure for strong queries. The All100 set has the Hard50 and New50 combined, and out of the 25 lowest performing t-init queries only 4 belong to the New50 set. Its 'area' measure behaves like Hard50 and drops after PRF.

The other four alternate queries defined by web-assistance all provide much better area measures for Hard50 as well as for All100 compared to t-init or t-prf. Thus, the web-based alternate representations by themselves are effective in improving the performance of these weak title topics. The MAP values are tabulated for information only and not considered for training purposes.

Title	**All100 (2003)**			**New50 (2003)**			**Hard50**		
Queries	**MAP**	**#0P10%**	**area**	**MAP**	**#0P10%**	**area**	**MAP**	**#0P10%**	**Area**
t-init	.1972	12	.0122	.2871	8	.0332	.1074	16	.0062
t-prf	.2414	17	.0108	.3496	8	.0439	.1332	26	.0036
Web-Assisted									
t-w_f-s	.2734	13	.0257	.3667	8	.0767	.1801	18	.0128
t-w_sf-s	.2672	11	.0229	.3467	10	.0424	.1877	12	.0153
t-w_s-s	.2618	11	.0186	.3720	8	.0511	.1516	14	.0112
t-pw_sf-s	.2662	13	.0193	.3474	12	.0405	.1850	14	.0122

Table 1: Training Results of Title Queries using TREC2003 Collection

Additional improvements for the Hard50 queries (and other query sets) were explored by combination of the above retrieval lists. We find good combination coefficients for two lists with area measure improvement as the objective by using grid search with steps of 0.1 or 0.05. Lists were combined recursively up to four. Using t-init or t-prf as basis for combination ensures that

the resultant retrieval list would not be too far off base. Our search is not exhaustive and the resultant combination is not optimal. On the other hand, we do not want to over-train on the Hard50 or the other sets either. Our 'title' submissions for TREC2004 consists of 4 runs (bolded) as follows: (to simplify reading, the number of snippets or full pages returned from web retrieval and the number of term output for alternate queries are suppressed):

pircRB04t1 – initial retrieval t-init based on the original TREC2004 titles;
pircRB04t2 – 2nd stage retrieval t-prf based on PRF;
pircRB04t3 – retrieval based on web-assistance that is a combination of the following retrieval lists with corresponding coefficients:
<**t-init**:0.15, **t-w_f-s**:0.4, **t-w_sf-s**:0.35, **t-w_s-s**:0.1>;
pircRB04t4 – retrieval based on web-assistance that is a combination of the following:
<**t-init**:0.2, **t-w_sf-s**:0.3, **t-w_s-s**:0.2, **t-pw_sf-s**:0.3>;
pircRB04t5 – un-submitted run that will be used later for 'td' retrieval. It is based on web-assistance that is a combination of the following:
<**t-prf**:0.1, **t-w_f-s**:0.4, **t-w_sf-s**:0.3, **t-pw_sf-s**:0.2>.

The first two submissions based on initial and PRF queries t-init and t-prf without web-assistance will not be competitive for robust evaluation, but they may be good for query ranking prediction (Section 4). pircRB04t3 and pircRB04t4 make use of different combinations of the initial retrieval with other alternate queries. Training with TREC2003 data shows that using initial retrieval as basis is more preferable. The Hard50 results for these combinations are shown in Table 3.

2.4 Queries from Description Section of TREC Topic

This section shows how alternate queries are formed when the longer description section of a TREC topic is used as query. These average out to 7.9 terms (with 2-word phrases) and about 70% of the queries are 6 terms or longer. Here we need term selection from the longer description section as was done last year (Laszlo et.al. 2004) in order to avoid cases of no web output. Term selection is difficult and could be often erroneous. This year we introduce a simple method of 'window rotation' to avoid salient term selection. We define a window of 5 terms and let it rotate through the description statement. For a description of m terms (m>5), there will be m such queries for web retrieval for each description. Returned web pages are ranked and selected based on their occurrence frequency in these m lists, and with frequency >= 2. When a query has <=5 terms, the 'window rotation method' defaults back to single retrieval. The resultant list of web items is used to define an alternate query. These two types of web retrieval are denoted as s (single) and r (rotation) in the last character of the query nomenclature.

d-init -- initial query from the description section of a topic processed with stop-word removal, Porter stemming, and some 2-word phrases.
d-prf(10,90) -- PRF query expansion with 10 top documents and 90 top terms from an initial retrieval.
d-pn_s-s(100,60) –this method assumes that the most content-bearing terms in a sentence are in phrases, followed by nouns. We consider three types of phrases in the order of: phrases identified by MINIPAR, phases containing nouns only (i.e. n gov n), and those also containing adjectives (adj gov n). We take all single words from the higher-ranking phrases, add to them all the nouns until the query contains six terms. We retrieve the top 100 snippets and select from them the 60 most frequent terms to form alternate query.
d-pnv_s-s(100,60) – this is similar to the previous except that verbs are also included with nouns for selection purposes.

d-w_sf-r(40,90) – content words retrieving 40 full pages and their snippets by window rotation, maximum 90-term alternate queries;
d-w_s-r(100,60) – content words retrieving 100 snippets by window rotation; maximum 60 term alternate query output;
d-pw_sf-r(40,90) – MINIPAR-identified 2-word phrases (and used as phrases for web retrieval) plus content words retrieving 40 web pages with their snippets by window rotation, maximum 90 term alternate queries.

Table 2 shows the results of training the alternate queries on the TREC2003 data. The initial and 2nd stage PRF retrieval results of these description queries (d-init, d-prf) provide better results than the corresponding title queries (t-init, t-prf) except in area measure of All100 for d-init vs. t-init, and #0P10% measure of Hard50 d-prf vs. t-prf. In Hard50, similar observations are true as for titles that area and #0P10% measures are worse for d-prf than for d-init. However, in contrast to title, the alternate queries for 'description' defined from the web are all inferior to d-init or d-prf in area measure for Hard50 except for one case (d-w_s-r). These alternate queries are weak because their web queries are not as precise as those derived from the titles. Apparently, human-generated short queries like the titles can solicit web texts that can form better alternate queries.

Description	**All100**			**New50**			**Hard50**		
Queries	**MAP**	**#0P10%**	**area**	**MAP**	**#0P10%**	**area**	**MAP**	**#0P10%**	**area**
d-init	.2342	9	.0121	.3503	4	.0638	.1182	14	.0063
d-prf	.2784	17	.0125	.4044	4	.0839	.1524	30	.0049
Web-Assisted									
d-pn_s-s	.2575	20	.0057	.3012	18	.0186	.1380	22	.0031
d-pnv_s-s	.2485	13	.0082	.3550	6	.0384	.1421	20	.0028
d-w_sf-r	.2147	18	.0070	.2852	16	.0172	.1442	20	.0035
d-w_s-r	.2497	11	.0090	.3626	8	.0343	.1368	14	.0055
d-pw_sf-r	.2543	14	.0076	.3737	10	.0249	.1373	18	.0044

Table 2: Training Results of Description Queries using TREC2003 Collection

After exploring various retrieval list combinations, our 'description' submission for TREC2004 also consists of 4 runs (bolded) as follows (with the Hard50 results for these combinations shown in Table 4):

pircRB04d1 – an un-submitted run based on initial retrieval using d-init.
pircRB04d2 – retrieval based on d-prf, after PRF;
pircRB04d3 – retrieval based on web-assistance that is a combination of the following:
<**d-prf**:0.3, **d-pn_s-s**:0.15, **d-w_sf-r**:0.25, **d-pw_sf-r**:0.3>;
pircRB04d4 – retrieval based on web-assistance that is a combination of the following:
<**d-prf**:0.35, **d-pnv_s-s**:0.2, **d-w_sf-r**:0.2, **d-w_s-r**:0.25>;
pircRB04d5 – same retrieval as pircRB04d3, but with different prediction for topic difficulty ranking;

2.5 Combining Title and Description Retrieval

Because of difficulties of selecting salient words, especially from the much longer narrative section of a TREC topic, we explored combination runs of Section 2.3 (Title queries) and Section 2.4 (Description queries) with the hope to further boost effectiveness. These are:

pircRB04td2 – retrieval list based on the combination of the following title and description retrieval lists: <**pircRB04d3**:0.5, **pircRB04t5**:0.5>;

pircRB04td3 – retrieval list based on the combination of the following title and description retrieval lists: <**pircRB04d3**:0.45, **pircRB04t3**:0.55>.

2.6 Results and Discussions

2.6.1 Title Queries

Table 3 shows the 'title' only runs named: pircRB04t1 to pircRB04t4. pircRB04t1 represents the simplest 'initial retrieval' run while pircRB04t2 refers to results of a second retrieval using pseudo-relevance feedback (PRF). The purpose of these runs is to see if we can predict their query difficulty ranking better than other more complicated runs that involve the web and retrieval combination. We know that their MAP and other evaluation measures will not be competitive with web-assisted runs. The web-assisted runs are named pircRB04t3 to pircRB04t5, but the last one was not submitted (indicated with a * in table) because of limits of submissions. Also shown in the table are the 'best' and 'median' results of each measure evaluated over all submitted 'title' runs.

Table 3 displays the set of 249 topics segmented into different sets, and their evaluation. In the following discussion, the results of 't3' (we will suppress the prefix pircRB04 from now on) will be used as an example to understand the evaluation values better. The 'All249' set contains the whole set of topics and differs from the 'Old200' (consisting of topics from TREC-6 to 8 and TREC2003) by the unseen 'New49' set (new for TREC2004). Set 'Hard50' (used in TREC2003) is a subset of the TREC6-8 topics within 'Old200'. The 'New49' query set has the high MAP value of 0.4008, which means this set has fewer low performing topics. If one looks at #0P10 which is the number of topics with zero precision at 10 documents retrieved, 'New49' has only 3, which when added to the 11 of 'Old200' gives a total of 14 for the full topic set 'All249'. The set 'Hard50' has more difficult topics not only because it has low MAP value of 0.1827, but also because its #0P10 value of 6 is 12% of its 50 topics compared to 11 or 5.5% for 'Old200'. The 'area' measure of 'Old200' 0.0333 is close to that of 'All249' 0.0376 since the differing set 'New49' probably has few contribution to the low performing topics. However, its 'area' measure .0333 is more than twice the .0158 value for 'Hard50', which needs some explanation. 'Old200' is a superset of 'Hard50' and it has quite an extra number of low performing topics to interleave into those from 'Hard50'. If we evaluate the area measure on the *same number* of topics, the value returned from 'Old200' would be smaller than that of 'Hard50'. However, the area definition is to use 25% of the lowest performing topics, which means counting 4 times as many topics in 'Old200' compared to 'Hard50'. This leads to the use of higher performing topics for the area value, and these contribute to a higher area value of .0333 for 'Old200'.

It is seen from Table 3 that between the two web-assisted runs 't3' and 't4', they achieved 12 of the total of 16 best precision values submitted for all 'title' runs. These are bolded. The 4 uncovered values are those of the 'Hard50' set. 't3' by itself has 8 of these 12 and seems to perform better than 't4'. The un-submitted run 't5' itself would have 7 measures (italicized) surpassing TREC2004 'top' values.

In general, the runs did very well for the 'New49' and the large sets 'Old200' and 'All249', but not as good for the #0P10 and area measures for the 'Hard50' set. For example, the 'Hard50' area for 't3' and 't4' are 0.0158, and 0.185 respectively. These are below the 'best' value of 0.0263. However, they are substantially better than the values of 0.0062 and 0.0036 achieved via initial retrieval 't1' or PRF 't2' respectively without web assistance. In general, our strategy of combining normal ad-hoc and alternate query retrieval lists works well, and training also

TITLE (TREC2004)			web-asist. 1432	web-asist. 2323	web-asist. 154351	PRF	Initial
	best	median	*pircRB04t5	**pircRB04t4**	**pircRB04t3**	**pircRB04t2**	**pircRB04t1**
Old200							
P10	**0.505**	0.437	*0.509*	0.494	**0.505**	0.439	0.407
map	**0.3165**	0.2468	*0.3191*	0.3129	**0.3165**	0.288	0.2431
#0P10	**10**	28	12	**10**	11	29	23
area	**0.0333**	0.0121	0.0329	0.0312	**0.0333**	0.0123	0.0121
New49							
P10	**0.549**	0.4245	0.5347	**0.549**	0.5449	0.4245	0.4143
map	**0.4019**	0.2856	0.4011	**0.4019**	0.4008	0.3408	0.2852
#0P10	**3**	6	*2*	3	**3**	10	6
area	**0.089**	0.0209	0.0668	0.0749	**0.0890**	0.0209	0.0259
Kendall's tau			.179	.121	.223	.121	.151
Hard50							
P10	**0.376**	0.28	*0.384*	0.362	0.374	0.28	0.268
map	**0.1942**	0.1152	*0.1943*	0.1695	0.1827	0.1332	0.1074
#0P10	**2**	11	6	5	6	13	8
area	**0.0263**	0.0059	0.0137	0.0185	0.0158	0.0036	0.0062
All249							
P10	**0.5129**	0.4361	*0.5141*	0.5048	**0.5129**	0.4361	0.4084
map	**0.3331**	0.2544	*0.3352*	0.3304	**0.3331**	0.2984	0.2514
#0P10	**13**	35	14	**13**	14	39	29
area	**0.0376**	0.0132	0.0363	0.033	**0.0376**	0.0132	0.0135
Kendall's tau	**0.623**	0.277	0.454	0.459	0.474	0.488	0.356

Table 3: Results of 'Title' only Runs

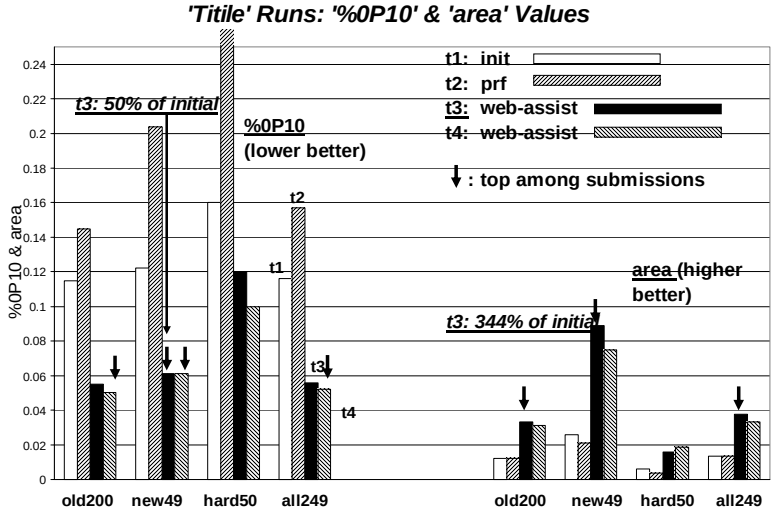

Fig.2: 'Title' Runs for Various Topic Subsets: '%0P10' and 'area' Values

DESCRIPTION (TREC2004)			web-asist. .3.15.25.3	web-asist. .35.2.2.25	web-asist. .3.15.25.3	PRF	initial
	best	**median**	**pircRB04d5**	**pircRB04d4**	**pircRB04d3**	**pircRB04d2**	*pircRB04d1
Old200							
P10	**0.508**	0.4535	0.504	0.5075	0.504	0.462	0.437
map	**0.3158**	0.2634	0.3139	**0.3158**	0.3139	0.299	0.2572
#0P10	**15**	30	15	17	**15**	32	20
area	**0.0303**	0.0092	0.0245	0.0234	0.0245	0.0112	0.0115
New49							
P10	**0.551**	0.4633	0.5408	0.5469	0.5408	0.4857	0.4673
map	**0.4074**	0.2992	0.4056	**0.4074**	0.4056	0.3717	0.3166
#0P10	**1**	5	2	**1**	2	5	1
area	**0.0739**	0.0245	0.0648	**0.0739**	0.0648	0.0404	0.0409
Kendall's tau			.330	.211	.175	.082	0.043
Hard50							
P10	**0.382**	0.316	0.372	**0.382**	0.372	0.322	0.306
map	**0.1635**	0.1328	0.1635	0.1622	**0.1635**	0.1524	0.1182
#0P10	**4**	9	5	6	5	15	7
area	**0.0205**	0.0071	0.0144	0.013	0.0144	0.0049	0.0063
All249							
P10	**0.5153**	0.4546	0.5112	**0.5153**	0.5112	0.4667	0.443
map	**0.3338**	0.2686	0.3319	**0.3338**	0.3319	0.3133	0.2689
#0P10	**17**	34	17	18	**17**	37	21
area	**0.0313**	0.0105	0.0273	0.0276	0.0273	0.0142	0.0141
Kendall's tau	**0.533**	0.318	0.318	**0.533**	0.514	0.503	0.52

Table 4: Results of 'Description' only Runs

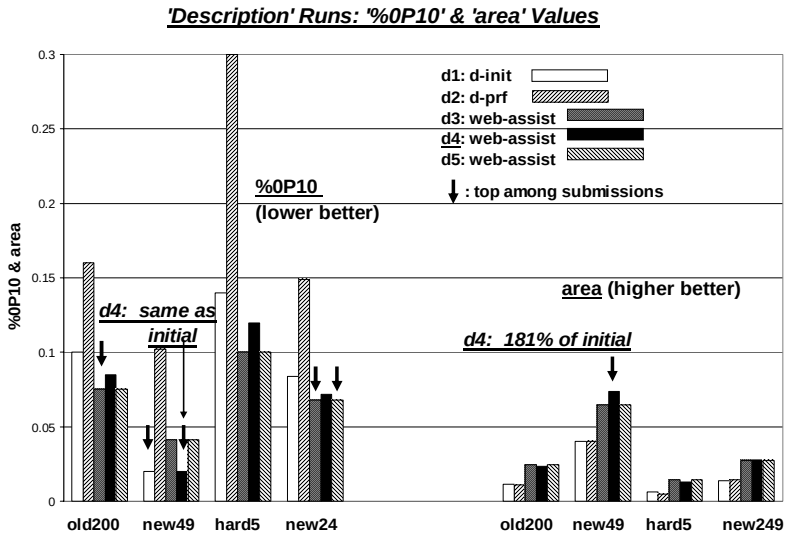

Fig.3: 'Description' Runs for Various Topic Subsets: '%0P10' and 'area' Values

generalizes nicely to the "New49' set. For example, run 't3' for 'New49' has all measures improve over 't1' initial and 't2' PRF runs. In particular, the 'best' area value of 0.089 improves over the 0.0209 of 't2' by 325% and over the 't1' value of 0.0259 by 243%. Fig.2 provides visual comparison of the two weak query measures for various topic subsets employing titles alone. It shows how poor the initial and PRF title queries (without web-assistance) perform in comparison.

2.6.2 Description Queries

Table 4 tabulates four runs we submitted using only the 'Description' section of a topic to define queries. These are similarly named as before: pircRB04d2 to pircRB04d5. In addition, 'd1' (leaving out the prefix pircRB04) is an un-submitted run that corresponds to results from an initial retrieval only. 'd2' is based on PRF, and 'd3', 'd4' are web-assisted runs. 'd5' retrieval is the same as 'd3' except that ranking of topic difficulty was done different. It is observed that 10 of the 16 best 'description' values were achieved between our 'd3' and 'd4' submissions. Of these two, 'd4' itself accounts for 7 of the 10 'best' values and generally has the better performance than 'd3' except for the 'area' measure of the 'Hard50' set. For these description runs, the area measures did not achieve the 'top' values except for 'New49' set, and covered all the best values for the measure #0P10 except for Hard50. As in title queries, our area measures for Hard50 set (0.0144 for 'd3' and 0.013 for 'd4') are below the best achieved of 0.0205. For the 'New49' query set, 'd4' has all evaluation measures equal or improve over those of 'd1' initial or 'd2' PRF retrievals. In particular, area measure of 0.0739 improves over 0.0404 or 0.0409 by more than 80%.

For the 'area' measure, web-assisted runs for 'description' are worse than for 'title' runs in all topic sets; but this is not true for the initial or PRF runs. For example for the New49, initial and PRF 'description' queries have area values of approximately .04, better than the corresponding 'title' values of about .02 to .025. However for web-assisted runs, these values for 'description' lie between .0648-.0749, worse than the corresponding 'title' values of .0668-.0890. It seems that the title words form effective web queries that produce effective alternate queries to supplement the original title retrieval, while this is not true starting from 'descriptions'. As before, Fig.3 shows that initial and PRF retrievals have lower effectiveness in '%0P10' and 'area' measures compared with web-assisted runs. The difference is not as wide as for 'title' queries.

2.6.3 Title + Description Queries

Two runs were submitted using a combination of selected 'title' and 'description' retrieval lists of the previous sections and their results are shown in Table 5. pircRB04td2 combines 'd3' (coefficient .5) and 't5' (.5), while td3 combines 'd3' (.45) with 't3' (.55). A third un-submitted run is td4 that combines 'd3' (.5) with 't4' (.5). Results show that 'td2' and 'td3' runs cover 10 of the 16 'best' values, and all measures are above median. The area values are less than the 'best' values except for the 'New49' topic set where 'td3' achieves the best value of 0.0917. Performing this combination between title and description retrieval lists generally brings additional small boost in the effectiveness measures. Fig.4 compares these 'td' runs with 't3' which is the best title run from Table 2, for various topic subsets. 'td2' and 'td3' perform equal or better than 't3' in all '%0P10' measures, while 'td4' has similar behavior for the 'area' results.

2.6.4 General Observations

Fig.5 summarizes the behavior of different topic lengths and topic sets for the four effectiveness measures: MAP, P10, #0P10% and area based on our submissions. It can be seen that the easiness of the query sets can be depicted as the following order: New49, All249, Old200 and

TITLE + DESCRIPTION (TREC2004)			web-assist. <.5d3,.5t4>	web-assist. <.45d3,.55t3>	web-assist. <.5d3,.5t5>
	best	**median**	*pircRB04td4	**pircRB04td3**	**pircRB04td2**
Old200					
P10	**0.5395**	0.451	0.539	**0.5395**	0.5385
map	**0.3429**	0.2667	0.3377	0.3422	**0.3429**
#0P10	**9**	23	12	10	**9**
area	**0.0573**	0.0129	0.041	0.0437	0.0437
New49					
P10	**0.551**	0.4449	0.5449	0.549	0.549
map	**0.4227**	0.2979	0.4193	0.42	**0.4227**
#0P10	**1**	5	2	3	3
area	**0.0917**	0.0265	*0.0958*	**0.0917**	0.0865
Kendall's tau			.119	.139	.201
Hard50					
P10	**0.402**	0.294	*0.41*	**0.402**	**0.402**
map	**0.1949**	0.126	0.1833	0.1918	**0.1949**
#0P10	**2**	9	5	4	3
area	**0.0457**	0.0072	0.0225	0.024	0.0222
All249					
P10	**0.5414**	0.4514	0.5402	**0.5414**	0.5406
map	**0.3586**	0.2755	0.3537	0.3575	**0.3586**
#0P10	**12**	28	14	13	**12**
area	**0.048**	0.0138	0.0456	0.0473	0.0468
Kendall's tau	**0.623**	0.266	.511	0.503	0.529

Table 5: Results of Combining Title & Description Runs

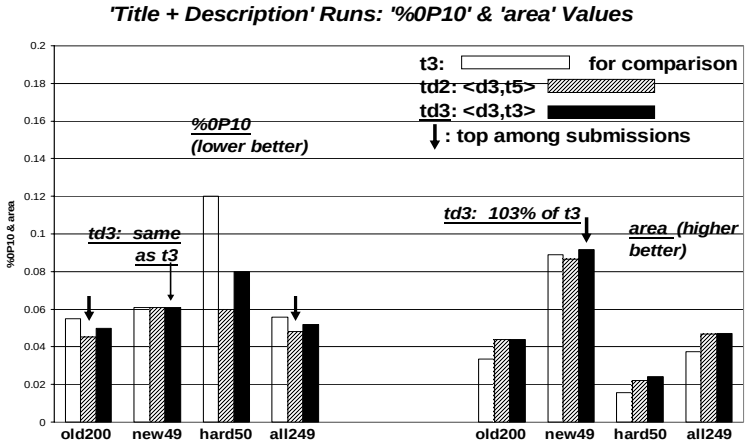

Fig.4: 'Title + Description' Runs for Various Topic Subsets: '%0P10' and 'area' Values

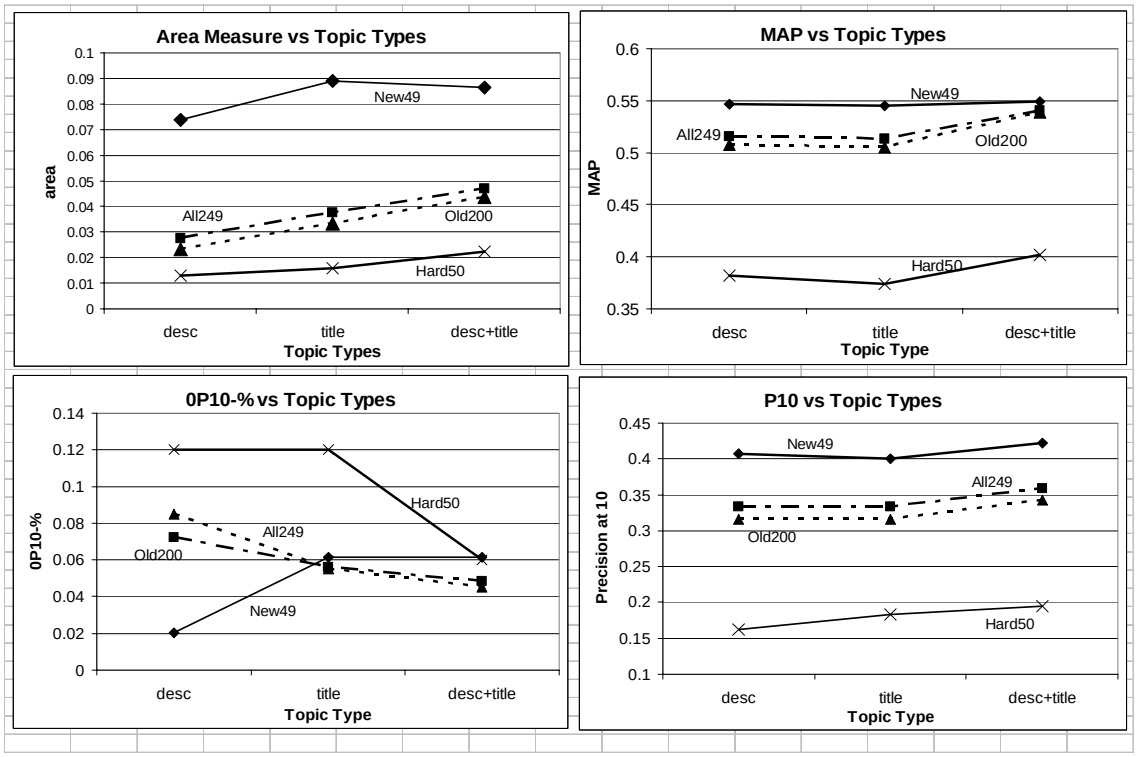

Fig.5: Area, 0P10-%, MAP and P10 vs Topic Type

Run ID	Median AP Best	(>/=/<)	Worst
pircRB04t1	6	108/10/131	0
pircRB04t2	21	171/4/74	0
pircRB04t3	28	206/4/39	0
pircRB04t4	16	208/4/37	0
pircRB04d2	21	154/8/87	0
pircRB04d3	19	189/18/42	0
pircRB04d4	14	199/6/44	0
pircRB04td2	5	215/2/32	0
pircRB04td3	3	215/2/32	0

Table 6: Comparing PIRCS All249 Results with Median

Hard50 (e.g. New49 plot is higher than others and (not quite so) lower for the 0P10% plot). It is also seen that, except for one or two exceptions, our performance with respect to topic sections is in the following order: 'td', 't', and 'd' (plots slant upwards to the right for all measures except 0P10% which slants downwards).

TREC2004 Robust track exercise shows that:

1. using the web as an all-domain thesaurus to improve topic representation is effective;
2. data fusion (combination of retrieval lists from web-assisted alternative queries) is

effective for improving retrieval, and for low performing topics in particular.

Compared to all submissions, our results perform very favorably. Table 6 shows the comparison using median AP and P10 values. For example, the web-assisted pircRB04t3 average precision has 206 topics better than median, 4 equal and 39 worse. 28 of the 206 have best average precision and none has worst. Tabulated MAP statistics for the increasing method index (e.g. from pircRB04t1 to pircRB04t4) also show the effectiveness of web-assistance for MAP values.

3 Robust Track – Predicting Topic Ranking by MAP

Over the years, TREC has provided 200 topics that have evaluation results with respect to the TREC-8 collections. From these topics one can form queries of different flavors (like topical content, wordings, specificity, etc.) and sizes (short or verbose, etc.). They span over a whole range of difficulties from MAP values of 0.0 to 1.0. A new task in TREC2004 Robust Track is to see whether the ranking of a given set of new queries can be predicted according to their retrieval difficulties based on training from the Old200 set.

From the task description, we decided that regression via the Old200 set for training and for parameter setting could be a viable approach. The 200 old topics will provide the features for characterizing the topics, and their MAP evaluation will provide the performance measure of difficulty. The major consideration at hand is to choose a set of reasonable features and to select the type of regression procedure that will do the prediction.

3.1 Choice of Features

Each query derived from a TREC topic section is represented by a set of terms. Experience from IR shows that terms with low document frequencies are good discriminators and therefore good for retrieval, and vice versa. We decided to use very simple features: locate the top 3 terms in each query and use their log D_k (log document frequency) plus their corresponding occurrence probability in the query (q_k/L_q). Here, D_k, q_k and L_q are respectively the document frequency of term k in the collection, the frequency of term k in query, and the length of a query (with minimum threshold set as 15 as used in PIRCS). These 6 basic features can be identified from a query description without any retrieval, and can be used to predict query ranking during initial retrieval, or for that matter, any subsequent retrievals.

After a retrieval is done, be it initial, PRF, etc., one will have a document list ranked by retrieval status value corresponding to a query under focus. We tried using coordinate matching measure as a feature to indicate how good or how bad a retrieval may be. When a query is short, like from the title section of a TREC topic, document having all the query terms could be a good indication of relevance, and vice versa. We count how many top-ranked documents n_x that have x=|q|, |q|-1, and |q|-2 words of the query, and use $\log(n_x)$ as additional features. |q| is the number of unique terms in a query. Together with |q|, these give a total of 10 features for our query characterization. These are specific to the initial query and its subsequent representation, and have been shown during n-fold training and validation to be mildly useful.

For this track, there are only 200 queries for training, which would limit the number of features one could employ. There can be more sophisticated features to use but due to time limitations, we have only considered these simple ones. Since there are different types of retrieval (initial, PRF, web-assisted, etc.), prediction of topic ranking can be done at these different stages. Coupled with the different query sizes (title, description, etc.), one can combine to provide a large number of retrievals and predictions. We have provided predictions only for the 10 runs

	#Features	Prediction for		
		Initial	PRF	Web-Assisted
Pre-retrieval - Query term properties	6: top-3 stems	t1		
After Initial Retrieval – Top-n documents	6: top-3 stems 1: query size 3: top-3 counts			
After PRF Retrieval – Top-n documents	6: top-3 stems 1: query size 3: top-3 counts		t2; d2	
After Web-Assist. Retrieval – Top-n documents	6: top-3 stems 1: query size 3: top-3 counts			t3, t4; d3, d4; td2, td3

Table 7: Features used for Topic Difficulty Prediction

discussed in Section 2, and summarized in Table 7. Prediction for pircRB04d5 made use of linear regression and will be discussed in Section 3.2.

3.2 Support Vector Regression

Support vector regression (SVR) is a relatively recent machine learning method for fitting sampled observations to a function f and using it to predict unseen data (Smola and Schölkopf 2004). In our task, given the training data $\{(x_1,y_1),.. (x_{200},y_{200})\}$, where x_i is a feature vector describing topic i ε Old200 and y_i its average precision, we want to predict the performance of y_j given features x_j (j ranges over the set New49) via the trained function. Like SVM for classification, it is nonlinear (and may be important for predicting MAP values via our simple features), scalable to large samples of high feature number efficiently (by being dependent on support vectors only and scalar product between data), and has been shown to be effective for various prediction problems (see for example Scholkopf, et.al. 1999).

We employed an implementation of SVR (called LIBSVM) that was developed at National Taiwan University and downloadable (Chang and Lin 2004). The simplest version, called epsilon-SVR or ε-SVR (Vapnik 1995) uses a radial-basis function as kernel although other functions such as polynomial can be defined. Several parameters need to be decided for the algorithm to work such as: epsilon ε (which defines a neighborhood around f where errors are considered tolerable), g (a parameter for the radial-basis function) and C (a cost value). For them, we have relied on the default values of $\varepsilon = 0.1$, $g = 1/\#\text{feature}$, and $C = 1.0$.

Training was done using the set Old200. Six basic term features were extracted from their queries for pircRB04t1. These are fed into LIBSVM and a model file (Mod200) was produced. For testing, the topic set All249 with their feature vectors together with Mod200 were used and LIBSVM produced predicted MAP values for each topic. Sorting the MAP values produced the ranking needed for submission. This procedure does not involve any retrieval, and would be very useful if it can predict topic difficulty.

The other runs all involved a set of ranked documents from a retrieval such as PRF or web-assistance. Only the top-ranked 200 retrieved documents were considered for feature extraction

for a total of 10 features as discussed in Section 3.1. The procedure for prediction is the same as for six features discussed in the previous paragraph.

3.3 Results of SVR Prediction

The official evaluation measure for this sub-task is Kendall's tau between the observed ranking and the predicted ranking of a topic set. tau = +1/-1 means complete agreement or disagreement in the two ranking results. For the title only experiments (Table 1), the simple 6-feature prediction of pircRB04t1 surprisingly gives a tau = 0.356 for the set All249. There is positive correlation between the lists of predicted and observed rankings. The other title runs all employ 10 features, and their tau values improve: varying between 0.459 (pircRB04t4) to 0.488 (pircRB04t2). When these tau values are converted to test statistics and consulted with the normal curve, they indicate that these rank correlations are all statistically highly significant. These high values are due to the fact that 200 of the 249 topics are used for training. When one considers the New49 set, which are unseen topics, their tau values dropped substantially to between 0.121 (pircRB04t4) to 0.223 (pircRB04t3).

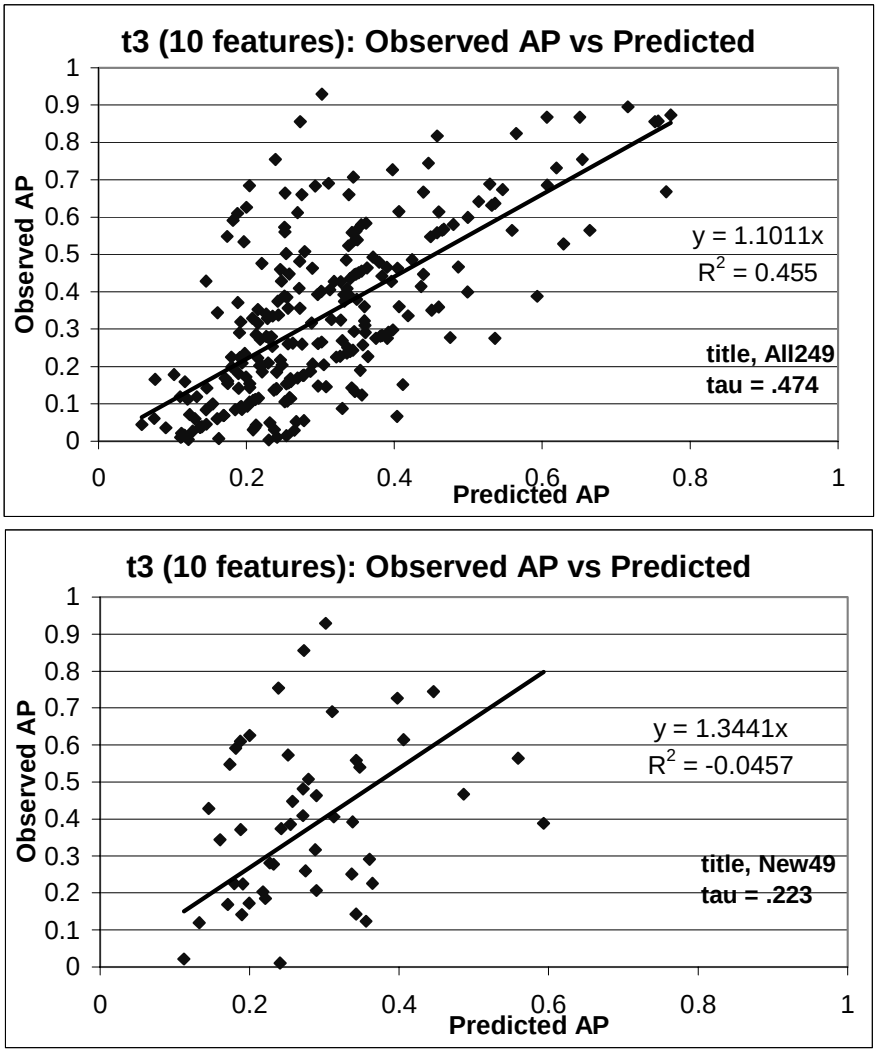

Fig.6: pircRB04t3 prediction of Average Precision for All249 and New49 Topics

For the description only runs, tau for All249 varies from 0.318 (pircRB04d5) to 0.533 (pircRB04d4); the latter is also the best value among all description submissions. Their New49 values vary from 0.082 (pircRB04d2) to 0.211 (pircRB04d4) and not as good as for titles. pircRB04d5 provides an exception with New49 tau value of 0.330. This prediction is discussed in Section 3.2. The two title + description runs have similar results as for the description runs.

The observation is that New49 topic prediction is better with the title topics, in particular pircRB04t3, using ε-SVR. The correlation is small. Fig.6 shows in greater details how prediction of individual topic precisions behaves for this run. The upper plot is for All249 set and the lower plot for the New49 set. The upper plot seems to show that the trend has been learnt reasonably well, although quite a few failed at the individual level. The lower plot shows that overall trend is not predicted (R^2 negative). Apart from the fact that regression does not predict average precision values well (e.g. lowest observed value is about 0.02 vs. predicted lowest value of 0.11), one can visually see that of the set of five worst ranked topics observed, prediction agrees with two of them. There is only one agreement for the set of five most effective observed. This information is not sufficiently accurate for one to employ specially tailored methods for the weakest or strongest topics.

3.4 Linear Regression

One of our submissions (pircRB04d5, which has the same retrieval as pircRB04d3 except for the topic difficulty prediction) was used to test linear regression as a prediction tool. For this run average query term weight (av-qtwt) of the PIRCS system was employed as feature because test of initial and PRF queries show that it has a small positive correlation with average precision. The query term weight qtwt involves the inverse collection term frequency and is defined in (Kwok 1995) as (N_w= number of tokens in the collection; F_k= collection frequency of term k):

$$qtwt = \log [q_k/(L_q-q_k)*(N_w-F_k)/F_k]$$

As discussed in Section 2.4, each of the topic in Old200 produces one standard TREC initial query and three alternate web-assisted queries, namely: **d-init**, **d-pn_s-s**, **d-w_sf-r** and **d-pw_sf-r.** These generate 8 sets of weights, 4 for initial retrieval and 4 for PRF defined from initial retrieval. Each of these 8 sets defines an average weight (av-qtwt) for each topic and they are employed as 8 features to predict the average precision value of pircRB04d5 using linear regression. This differs from the feature choice of Section 3.1 in that the same attribute from many query types of the same topic are used, rather than several attributes from the same query. This appears quite costly. However, all the weight files are generated as a by-product of the procedure that results in the combination run pircRB04d5 of 4 retrievals. These 4 retrievals can be speeded up substantially via parallel hardware and processing if available.

As displayed in Table 4, the result of this run is surprising. Its Kendall's tau value for All249 is 0.318 which is not too low compared with our other submissions. However, for the unseen topics New49, its prediction has tau = 0.330, substantially better than other runs. This appears to suggest that the diverse query types of the same topic may contribute clues for difficulty prediction. This has to be studied in greater details with more experimental observations.

4 Conclusion

For a second year in a row, we have demonstrated that our approach of exploiting the web (probing the web to return relevant/related output to define alternate queries for a given query, and combine their retrieval lists) to enhance ad hoc retrieval is viable and effective. This method

can improve ‘area’ measure over 300% and reduce the number of ‘0P10’ queries by 50% compared to initial retrieval using short (title) queries of the unseen New49 set. Similarly, it improves over 180% in ‘area’ while maintaining the same number of ‘0P10’ queries when medium length (description) queries were employed. These medium ‘description’ queries do not provide as effective alternate queries as the short ‘titles’ because they are longer and requires term selection to probe the web. Appropriate term selection is difficult. We introduced a window rotation method that does not rely on term selection – it is stable and effective but time consuming. An important topic of research is how to do term selection and compose web queries from given ad hoc queries when these are medium to long in length.

We have experimented with support vector regression to predict new query effectiveness from old data using some simple features. Results however are not sufficiently accurate to be used to do individual query tailoring. Apart from the fact that the choice of features might be improved for prediction, the number of old data (200 queries) for training may also be not sufficient.

Acknowledgment

This work was partially supported by a U.S. Govt. DST/ATP contract 2003*H532600*000.

References

Chang, C-C and Lin C-J (2004). LIBSVM: a Library for Support Vector Machines. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Dorr, B., Zajic, D & Schwartz, R (2003). *Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation.* Proceedings of HLT-NAACL 2003 Text Summarization Workshop, Paper W03-0501.

Grunfeld, L., Kwok, K.L., Dinstl, N. & Deng, P. (2004). TREC 2003 Robust, HARD and QA Track Experiments using PIRCS. In: Information Technology: The Twelfth Text REtrieval Conference, TREC 2003. E.M. Voorhees & L.P. Buckland, Editors. NIST Special Publication 500-255, US GPO: Washington, DC. pp.510-521.

Kwok, K.L (1995). A network approach to probabilistic information retrieval. ACM TOIS 13:324-353.

Lin, D. (1994). PRINCIPAR – an efficient, broad-coverage, principle-based parser. *Proc of COLING-94*. pp.482-488

Scholkopf, B., Burges, C.J.C & Smola, A.J., editors (1999). *Advances in Kernel Methods – Support Vector Learing.* MIT Press.

Smola, A.J. and Schölkopf, B. (2004). A tutorial on Support Vector Regression. In *Statistics and Computing* 14: 199-222.

Vapnik V. (1995). *The Nature of Statistical Learning Theory*. Springer, 1995.

Voorhees, E.M. (2004). Overview of the TREC 2003 Robust Retrieval Track. In: Information Technology: The Twelfth Text REtrieval Conference, TREC 2003. E.M. Voorhees & L.P. Buckland, Editors. NIST Special Publication 500-255, US GPO: Washington, DC. pp.69-77.