University of Essex at the TREC 2011 Session Track

M-Dyaa Albakour[1], Udo Kruschwitz[1], Nikolaos Nanas[2]
Brendan Neville[1], Deirdre Lungely[1], Maria Fasli[1]
[1] School of Computer Science and Electronic Engineering,University of Essex, Colchester, UK
[2] Centre for Research and Technology - Thessaly (CE.RE.TE.TH), Volos, Greece
{ malbak | udo | bneville | dmlung | mfasli } @essex.ac.uk, n.nanas@cereteth.gr

ABSTRACT

This paper provides an overview of the experiments we carried out at the TREC 2011 Session Track. We propose two different approaches to tackle the task introduced this year. The first one relies on a biologically inspired adaptive model for information filtering to build a user profile of multiple topics of interests throughout the session. The learnt profile is then exploited in the retrieval process. The second approach is an extension of our anchor log technique we proposed in the previous year. We use the anchor logs to simulate queries in order to derive query expansions that are relevant to user information needs throughout the session.

1. INTRODUCTION

The Session Track was introduced at the Text REtrieval Conference (TREC) 2010. The Session Track aims to evaluate the ability of search engines to utilise previous user interactions in order to provide better results for subsequent queries in a user session and therefore 'point the way' to what the user is actually looking for. Last year participants were given query sessions containing only two queries and no interaction data. This year the session track provided more interactive data to the participants. Query sessions were collected from real users and contained a variable number of queries and interaction data such as the documents displayed to the user, the clicked documented and dwelling times.

This year the session track was another opportunity to explore different adaptive modelling techniques to improve retrieval over query sessions. The wider context is the AutoAdapt project[1] which looks at automatically building and adapting domain models from the users' search and browsing behaviour (using query logs). These domain models are used to assist users to find information by suggesting query modification or browsing suggestions in their search.

We submitted three runs based on two adaptive models. The first of these is Nootropia, a biologically inspired adaptive modelling technique that has been successfully applied to Information Filtering (IF) problems such as news recommendation [10]. We used Nootropia to adapt a profile of information needs throughout the session and exploited the profile in the retrieval process for the last query. Two different strategies were used to build the adapted profile. The first models what the user is interested in and conversely the second models what the user is not interested in.

The second approach is an extension of our anchor log expansion technique we proposed in the previous year. We exploit the anchor logs to derive query expansions that are relevant to user information needs throughout the session. Anchor text has shown to be effective for a variety of information retrieval tasks. This includes ad-hoc search and the diversity task [9], [3]. Anchor text can be considered as a replacement to user queries as often web authors use similar labels to describe web pages to those used by searchers to find them [5]. Moreover, Dang and Croft have recently shown how anchor text can be used to simulate user sessions. They have considered all the anchor text pointing to the same document as queries in the same user session [4]. We derived query expansions from the anchor logs by either using anchor logs as simulation of query logs to derive related terms or phrases that represent the user information needs in the session or by using an implicit relevance feedback model that exploits the anchor text of the displayed or clicked documents throughout the session.

The rest of the paper is structured as follows. In Section 2 we give a brief description of the task introduced this year. We describe the dataset and the resources used in our runs in Section 3. We explain the experiments and the runs submitted to TREC in Section 4. The results of those runs are then discussed in Section 5. Finally we give a brief conclusion in Section 6

2. THE TASK

The main difference in this year's task compared to the last year is that more interactive data is provided to the participants.

Participants have been provided with a set of query sessions. Each session consists of the current query q_m and the query session prior to the current query:

(a) the set of past queries in the session q_1, q_2, ..., q_{m-1}.

(b) the ranked list of URLs for each past query,

(c) the set of clicked URLs/snippets and the time spent by the user reading the corresponding to each clicked url web page.

[1] `http://autoadaptproject.org`

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
TREC '11 Gaithersburg, USA
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$10.00.

Participants then run their retrieval system over the current query,

- ignoring the session prior to this query ($RL1$).
- considering only the item (a) above, i.e. the queries prior to the current query ($RL2$).
- considering only the items (a) and (b) above, i.e. the queries prior to the current along with the ranked lists of URLs and the corresponding web pages ($RL3$).
- considering only the items (a), (b) and (c) above, i.e the queries prior to the current, the ranked lists of URLs and the corresponding web pages and the clicked URLs and the time spent on the corresponding web pages ($RL4$).

3. EXPERIMENTAL SETUP

The ClueWeb09 dataset[2] is a web crawl of more than a billion pages that has been used in last year's Web track. The ClueWeb09 category B dataset is a subset of the larger ClueWeb09 crawl and it consists of 50 million English pages. In this year's task participants were permitted to use either one of the two datasets. An existing Indri[3] index of the ClueWeb09 dataset is already available and searchable via a public web service[4]. The web service would enable us to issue queries and retrieve the top documents returned by the search engine, thus removing the burden of indexing the data internally. The Indri search engine uses language modelling probabilities and supports query expansion.

The anchor log for the dataset has been processed and made publicly available by the University of Twente[5]. Each line in the log represents a document in the collection with all the anchor text pointing to the document [6]. We used the anchor log file of the ClueWeb09 Category B dataset. This file contains 43 million lines and thus contains anchor text for about 87% of the documents. Each line is tab separated and consists of the document TREC identifier, its url and all the anchor text pointing to that document.

Figure 1 shows a sample line in the anchor log file. We add quotation marks to group anchor text fields for illustration purposes.

In the next sections describing our runs, we will use the following terminology. For a query q consisting of a number of terms qt_i, our reference search engine (The Indri search engine) would return a ranked list of documents using the query likelihood model from the ClueWeb09 category B dataset:

$D_q\langle d_{q,1}, d_{q,2}, ..., d_{q,n}\rangle$ where $d_{q,i}$ refers to the document ranked i for the query q based on the reference search engine's standard ranking function.

4. RUNS

As a group participating in the session track we can submit three different runs to TREC. Each run consists of the four lists $RL1$, $RL2$, $RL3$, and $RL4$.

[2] http://boston.lti.cs.cmu.edu/Data/clueweb09/

[3] http://lemurproject.org/indri.php

[4] http://boston.lti.cs.cmu.edu:8085/clueweb09/search/cataenglish/lemur.cgi

[5] http://wwwhome.cs.utwente.nl/ hiemstra/2010/anchor-text-for-clueweb09-category-a.html

In all the runs we will generate $RL1$ by simply submitting a preprocessed version of the current query q'_m to the Indri index, i.e. $RL1$ will be equivalent to $D_{q'_m}$.

The current query q_m was processed to produce q'_m following these steps:

1. Removing the following punctuation marks ( ) , ?
2. Removing stop words from a common list of English stop words that do not fall within quoted text
3. Replacing quotes with the corresponding Indri syntax #1(<quoted text>) e.g. *"event planning"* becomes *#1(event planning)*
4. Replacing site specification with the corresponding Indri format, e.g. "female winemakers site:.com.au" becomes "female winemakers com.url au.url"

The maximum number of returned documents in the list is limited to 1000. We also used the Waterloo Spam Rankings[6] for the ClueWeb09 dataset to filter the spam documents from the returned ranked lists. We consider documents with scores of 70% or less as spam which is recommended by the creators of those rankings [2].

Generating $RL2$ would be a similar task of generating $RL3$ in the previous year's task [8]. In our runs we will generate $RL2$ in different ways.

The major challenge resides in generating $RL3$ and $RL4$. In the following we propose a couple of methods to generate those lists.

4.1 Nootropia to model query sessions

Nootropia is a biologically inspired Information Filtering system that has been used successfully in news recommendation [10].

With Nootropia, information needs are represented as a weighted and ordered network that may represent an individual's multiple topics of interest. This network profile presentation can evaluate the relevance of an information item to the user interests based on a directed spreading activation model. Nodes (features) in the network that appear in the item are activated and subsequently disseminate part of their initial energy to nodes with larger weight. At the end of this feedforward dissemination process, a single relevance score can be calculated as the weighted sum of the final energies of activated terms.

The proposed process of using Nootropia to generate $RL3$ and $RL4$ is illustrated in Figure 2. Throughout the session we take the documents that have been displayed to the user by the search engine to create a Nootropia model. For the current query q_m, we submit the preprocessed version q'_m, as explained before, to the search engine and then we use the profile built so far in a form of a Nootropia network to give a score to the returned documents reflecting how much they match the profile. The scores can be used to re-rank the documents. We will use two routes to re-rank the documents using the Nootropia profile and therefore we propose two different runs:

- **Nootropia Positive (essexNooPos)**: To generate $RL3$, we can use all the displayed documents each time to build a Nootropia profile and update it. Here

[6] http://durum0.uwaterloo.ca/clueweb09spam/

clueweb09-en0000-23-00060 *http://001yourtranslationservice.com/dtp/* 'website design' 'DTP and Web Design' 'Samples' 'programmers' 'desktop publishing' 'DTP pages' 'DTP samples' 'DTP and Web Design Samples' 'DTP and Web Design Samples' 'DTP and Web Design Samples' 'DTP and Webpage Samples' 'DTP' http://001yourtranslationservice.com/dtp/

Figure 1: A sample line in the anchor log file

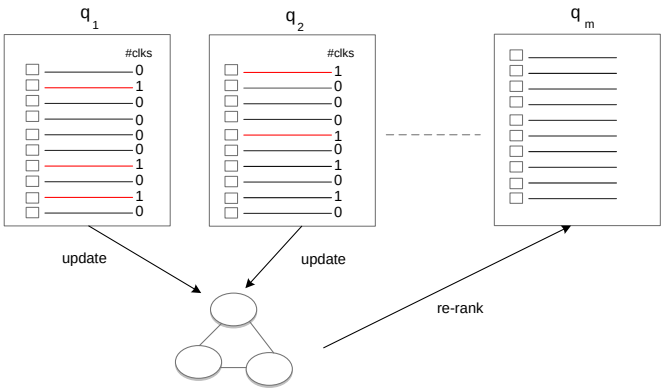

Figure 2: Illustration of the process of applying Nootropia

we take an "optimistic" approach and consider displayed documents as useful documents to the user. The documents that match this profile should be promoted. Therefore we rank the documents returned by the search engine for the current query by their Nootropia score in an *descending* order and that would be our $RL3$.

Let N^a be the Nootropia profile for all displayed documents to the user.
Let $S_{N^a}(d)$ be the matching score of a document d in the profile N^a.

$RL3 = \langle d_1, d_2, .., d_n \rangle; \forall i : d_i \in D_{q'_m} \wedge S_{N^a}(d_i) \geq S_{N^a}(d_{i+1})$

For $RL4$, we take only the clicked documents to build a profile of user interests unlike the previous profile.

Let N^+ be the Nootropia profile for all clicked documents by the user.
Let $S_{N^+}(d)$ be the matching score of a document d in the profile N^+.

$RL4 = \langle d_1, d_2, .., d_n \rangle; \forall i : d_i \in D_{q'_m} \wedge S_{N^+}(d_i) \geq S_{N^+}(d_{i+1})$

- **Nootropia Negative (essexNooNeg)**: Here we take a "pessimistic" approach and consider the displayed documents not useful to the user as they had to reformulate the query and therefore any document that matches this profile should be penalised. Therefore we rank the documents returned by the search engine for the current query by their Nootropia score in an *ascending* order and that would be our $RL3$.

  Let N^a be the Nootropia profile for all displayed documents to the user.
  Let $S_{N^a}(d)$ be the matching score of a document d in the profile N^a.

  $RL3 = \langle d_1, d_2, .., d_n \rangle; \forall i : d_i \in D_{q'_m} \wedge S_{N^a}(d_i) \leq S_{N^a}(d_{i+1})$

  As for $RL4$, we take only the documents not clicked on to build a profile of documents that the user is not interested in. The documents that match this profile should be penalised. Therefore, we rank the documents returned by the search engine for the current query by their Nootropia score in an *ascending* order and that would be our $RL4$.

  Let N^- be the Nootropia profile for all abandoned documents by the user.
  Let $S_{N^-}(d)$ be the matching score of a document d in the profile N^-.

  $RL4 = \langle d_1, d_2, .., d_n \rangle; \forall i : d_i \in D_{q'_m} \wedge S_{N^-}(d_i) \leq S_{N^-}(d_{i+1})$

To update the profile, we should feed the document or a representation of the document to the Nootropia model. For this we extracted all the nouns and noun phrases in the snippet of the document. For the detection of noun phrases we look for particular patterns of sequence of part-of-speech tags based on the algorithm for the detection of terminological terms as described in [7].

In both runs, to generate $RL2$ we use a similar approach to the one used in our baseline system last year (i.e. 'essex1') where we used the previous history of the user by simply submitting both queries in the session. Therefore, to generate $RL2$ for both Nootropia runs we simply submit a query which consist of the first query and the last query in the session q_1, q_m.

4.2 Expansion with anchor text (essexAnchor)

The second approach we propose here can be seen as an extension to our anchor log technique proposed in the previous year [1].

To generate $RL2$ we use a similar approach to the one used in our anchor log system last year ('essex3'). We consider

the first query and the last query in the session q_1, q_m and use the same method described in our last year's paper [1] to generate query expansions from the anchor log to the reformulated query q_m.

For $RL3$ and $RL4$, we use a relevance feedback approach by expanding the query with the anchor text of displayed or clicked documents respectively.

$$\begin{array}{c} \#\ \text{combine}(\\ 0.7\ \#\ q_m\ 0.3\ \#\ \text{combine}(\ w_1 \#e_1\ w_2 \#e_2\ ..\ w_{10} \#e_{10}) \\ ) \end{array}$$

w_i is the normalised frequency of the anchor text across the documents.

In Table 1 we give an example of the expansions extracted to generate queries for $RL3$, $RL4$ for session 63.

5. RESULTS

Tables 2 and 3 summarise the results for our three runs as well as the maximum and median for all the session track participants. This year NIST assessors provided relevance assessments on two different criteria thus the results in Table 2 take into account all the subtopics, whereas the results in Table 3 only reflect the user's last subtopic.

Each of the table columns represent the normalised discounted cumulative gain (nDCG) for each result list submitted, despite receiving a variety of relevance metrics we will focus our results analysis on the nDCG results, as they relate to the established metrics from last year's competition. The tables are split in two, the top half using the nDCG metric for all the documents submitted and the bottom showing the nDCG score using only the first ten returned documents (nDCG@10).

The arrows in the RL2, RL3 and RL4 columns represent the relative improvement or decline in the nDCG scores between the results lists for a given system (row), with a double ($\Uparrow$) arrow indicating that a two tailed t-test has supported the result as significant. For instance an upward arrow ($\uparrow$) in the RL2 column indicates that RL2 improves on RL1, and the first and second arrows in the RL3 column compare RL3 to RL1 and RL3 to RL2 respectively. These results are also charted without significance test data in the graphs in Figure 3. The following subsections outline the results obtained for each system, in cases where we do not explicitly refer to a result as significant it can be assumed that the comparison has returned a t-test value $p > 0.05$.

5.1 Expansion with Anchor Text

Taking all the subtopics (Table 2) we see that the retrieval performance of RL2 is worse than RL1 when looking at all retrieved documents, but conversely the nDCG@10 score does improve i.e. the top results improve. This is in line with last year's results where expansions using association rules from query logs simulated from anchor logs improved performance. However, using only the last subtopic RL2 performs worse than RL1 on both nDCG and nDCG@10.

Comparing RL3 to RL1 demonstrates significant improvement in all cases i.e. when using the last subtopic, all the subtopics and against nDCG and nDCG@10. Therefore the implicit relevance feedback from anchor text of retrieved documents does improve the retrieval performance.

RL4 also shows improvement relative to RL1 in all cases, however it is only significant when using all subtopics. RL3 is marginally better than RL4 in all cases though in general the difference is so small as to consider them on par. Thus using the anchor text only from clicked documents to expand queries does not appear to improve performance.

Relevance scores for our non-baseline submissions (RL3 and RL4) are between median and maximum of the composite scores for the track and when using nDCG@10, their performance is very close to maximum and well above median results reported. In fact 'essexAnchor' was the top performing system among all the participants when it comes to nDCG@10 for RL3 using all subtopics. The anchor text approach performs better when assessed over all the subtopics than when measured against only the last topic, which is not surprising as this method tries to model user interest throughout the session and does not only target a specific subtopic of a query. Note that this correlates with the overall maximum and median scores.

5.2 Nootropia

When assessing all subtopics RL2 is significantly better than RL1 for both nDCG and nDCG@10. RL2 combines the user's first and last query and it was used as our baseline in last year's session track The results here are in line with last year where this baseline has shown to be capable of improving the retrieval performance over an adhoc system. However considering only the last subtopic, RL2 improves on RL1 for nDCG but is worse for nDCG@10.

Comparing the relevance scores for RL3 and RL4 to RL1 shows performance declines when using Nootropia Positive using either evaluation strategies and in some case the difference is significant. While RL3 and RL4 both show a similar downward trend in comparison with RL1, only considering the clicked documents in RL4 does improve on RL3. Like our anchor test method the retrieval performance using all sub-topics is better than just using the final subtopic, this follows the general trend seen in the summary results provided by NIST i.e. the maximum and median scores illustrated in tables 2 and 3.

The results for the Nootropia Negative run (RL3 and RL4), show that the "pessimistic" approach performs significantly worse than our baselines (RL1 and RL2) across all metrics. In addition using the click data (RL4) results in a further decline in retrieval performance. Comparing the results for Nootropia Negative (both RL3 and RL4) to Nootropia Positive we see that Nootropia Positive performs significantly better.

6. CONCLUSION

The session track in TREC 2011 provided a framework for testing retrieval systems over query sessions where previous queries and interactions with the search engine can be utilised in the retrieval process.

As in the previous year, we show that the anchor log of the document collection is a valuable resource that can be useful in the session retrieval problem. In the last year's task, previous queries in the session were useful to derive query expansions for reformulated queries from the anchor log to improve the retrieval performance. This finding was supported by the results this year, and in addition to that, with interaction data the retrieval performance is further enhanced by using the relevance feedback from the anchor text of retrieved documents. This approach achieved the best retrieval performance for RL3 among all participating systems in the track.

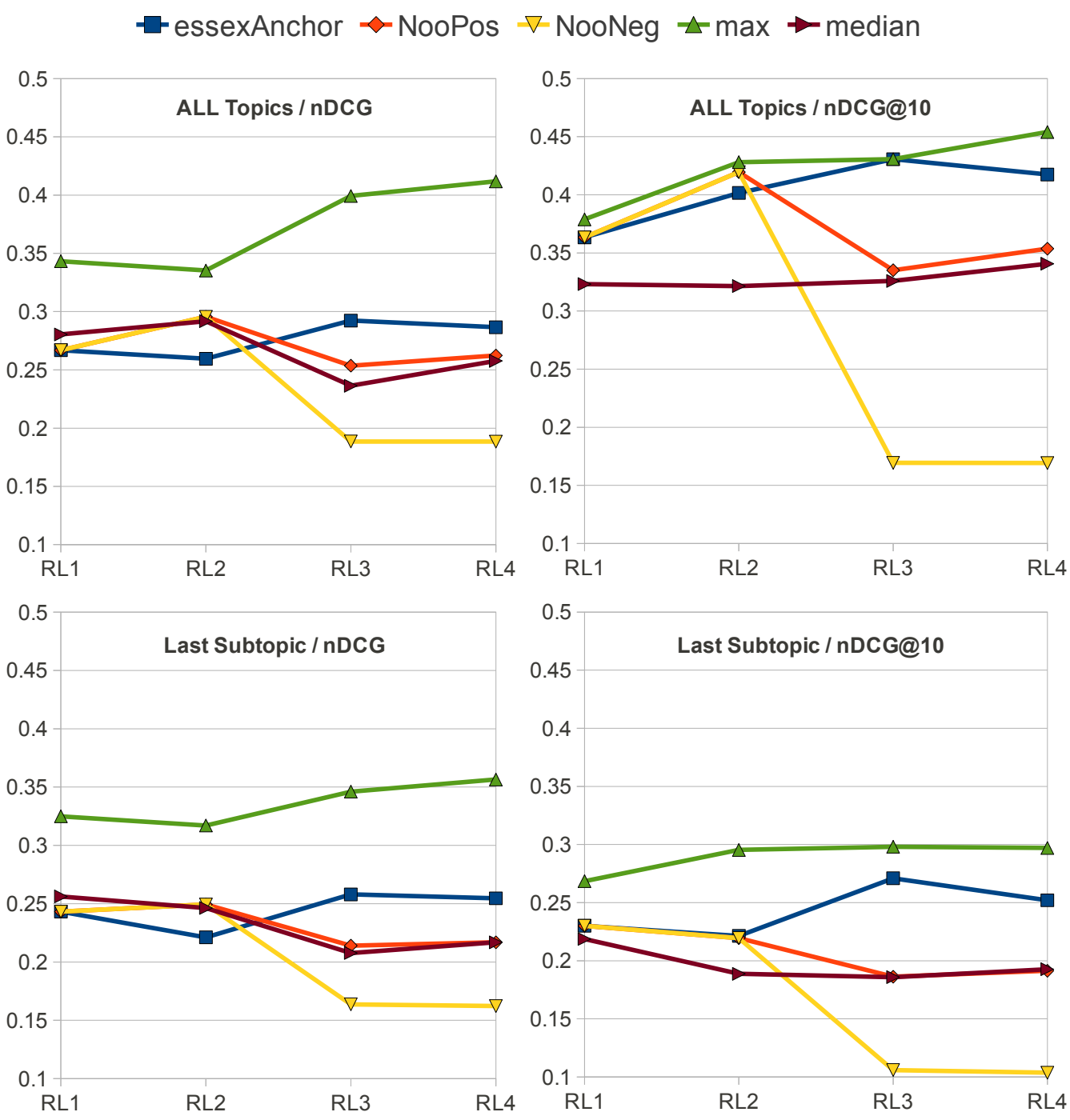

Figure 3: Graphs showing nDCG and nDCG@10 results

Table 1: Example of expansion terms and phrases extracted for session 63: FISA → judges on fisa court → 1990 FISA wiretap applications → judges FISA court → judges FISA court 2005

	Expansion:weight
RL3	united states foreign intelligence surveillance court:0.31 foreign intelligence surveillance act:0.23 united states foreign intelligence surveillance court of review:0.10 fisc:0.08 fisa:0.07 section summary of the usa patriot act title ii:0.06 usa patriot act title ii:0.05 ii:0.03 title ii enhanced surveillance procedures:0.03 enhanced surveillance procedures:0.03
RL4	foreign intelligence surveillance act:0.78 fisa:0.16 article:0.02 foreign intelligence surveillance act fisa:0.02 http en wikipedia org wiki foreign intelligence surveillance act:0.02

Table 2: nDCG values when assessing over all subtopics; the arrows indicate improvement(↑) or decline (↓) against the previous results lists, the first arrow in a cell relates to RL1, the second arrow to RL2 and so on. Double arrows (⇑ / ⇓) indicates the comparison is statistically significant returning a two tailed t-test $value \leqslant 0.05$. The figure in bold is the top obtained score for that measure in session track 2011.

System	RL1.nDCG	RL2.nDCG	RL3.nDCG	RL4.nDCG
max	0.3433	0.3353	0.3993	0.4112
median	0.2804	0.2918	0.2363	0.2577
essexAnchor	0.2669	↓ 0.2595	⇑ ⇑0.2923	⇑ ⇑ ↓ 0.2866
essexNooPos	0.2669	⇑ 0.2956	↓ ⇓ 0.2536	↓ ⇓ ↑ 0.2623
essexNooNeg	0.2669	⇑ 0.2956	⇓ ⇓ 0.1885	⇓ ⇓ ↓ 0.1884
	RL1.nDCG@10	RL2.nDCG@10	RL3.nDCG@10	RL4.nDCG@10
max	0.3789	0.4281	0.4307	0.4540
median	0.3232	0.3215	0.3259	0.3407
essexAnchor	0.3634	↑ 0.4016	⇑ ↑ **0.4307**	⇑ ↑ ↓ 0.4175
essexNooPos	0.3634	⇑ 0.4195	↓ ⇓ 0.3351	↓ ⇓ ↓ 0.3536
essexNooNeg	0.3634	⇑ 0.4195	⇓ ⇓ 0.1694	⇓ ⇓ ↓ 0.1692

Table 3: nDCG values when assessing the last subtopic; the arrows indicate improvement(↑) or decline (↓) against the previous results lists, the first arrow in a cell relates to RL1, the second arrow to RL2 and so on. Double arrows (⇑ / ⇓) indicates the comparison is statistically significant returning a two tailed t-test $value \leqslant 0.05$.

System	RL1.nDCG	RL2.nDCG	RL3.nDCG	RL4.nDCG
max	0.3249	0.3170	0.3461	0.3565
median	0.2562	0.2463	0.2077	0.2169
essexAnchor	0.2432	↓ 0.2211	⇑ ⇑ 0.2580	↑ ⇑ ↓ 0.2546
essexNooPos	0.2432	↑ 0.2493	⇓ ⇓ 0.2140	⇓ ⇓ ↑ 0.2169
essexNooNeg	0.2432	↑ 0.2493	⇓ ⇓ 0.1637	⇓ ⇓ ↓ 0.1622
System	RL1.nDCG@10	RL2.nDCG@10	RL3.nDCG@10	RL4.nDCG@10
max	0.2685	0.2954	0.2981	0.2971
median	0.2187	0.1888	0.1859	0.1927
essexAnchor	0.2301	↓ 0.2214	⇑ ⇑ 0.2710	↑ ↑ ↓ 0.2520
essexNooPos	0.2301	↓ 0.2195	↓ ↓ 0.1863	↓ ↓ ↑ 0.1914
essexNooNeg	0.2301	↓ 0.2195	⇓ ⇓ 0.1059	⇓ ⇓ ↓ 0.1037

We also show how an adaptive IF system can be employed for the session retrieval problem using two different strategies. For the proposed setup, no improvement was observed using either strategies. However with the reusable dataset and relevance judgements provided by NIST we will conduct more experiments where we develop the approach presented here in applying this adaptive model.

Acknowledgements

This research is part of the AutoAdapt research project. AutoAdapt is funded by EPSRC grants EP/F035357/1 and EP/F035705/1.

7. REFERENCES

[1] M.-D. Albakour, U. Kruschwitz, J. Niu, and M. Fasli. University of Essex at the TREC 2010 Session Track. In *Proceedings of the 19th Text REtrieval Conference (TREC'10)*, 2011.

[2] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, 2010.

[3] N. Craswell, D. Fetterly, M. Najork, S. Robertson, and E. Yilmaz. Microsoft research at trec 2009: Web and relevance feedback tracks. In *Proceedings of the 18th Text REtrieval Conference (TREC)*. NIST, 2009.

[4] V. Dang and B. W. Croft. Query reformulation using anchor text. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 41–50, New York, NY, USA, 2010. ACM.

[5] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 459–460, New York, NY, USA, 2003. ACM.

[6] D. Hiemstra and C. Hauff. Mirex: Mapreduce information retrieval experiments. Technical Report TR-CTIT-10-15, Centre for Telematics and Information Technology University of Twente, Enschede, April 2010.

[7] J. S. Justeson and S. M. Katz. Technical terminology some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

[8] E. Kanoulas, P. Clough, B. Carterette, and M. Sanderson. Session Track at TREC 2010. In *Proceedings of the Workshop on the Automated Evaluation of Interactive Information Retrieval in conjunction with the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010*, Geneva, Switzerland, 2010.

[9] M. Koolen and J. Kamps. The importance of anchor text for ad hoc search revisited. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129, New York, NY, USA, 2010. ACM.

[10] N. Nanas and A. N. D. Roeck. Autopoiesis, the immune system, and adaptive information filtering. *Natural Computing*, 8(2):387–427, 2009.