# QUT_Para at TREC 2012 Web Track: Word Associations for Retrieving Web Documents

Mike Symonds[1], Guido Zuccon[2], Bevan Koopman[1,2], Peter Bruza[1]

michael.symonds@qut.edu.au, guido.zuccon@csiro.au,
bevan.koopman@csiro.au, p.bruza@qut.edu.au

[1] School of Information Systems, Queensland University of Technology
[2] Australian e-Health Research Centre, CSIRO

Brisbane, Australia

**Abstract**

Many existing information retrieval models do not explicitly take into account information about word associations. Our approach makes use of first and second order relationships found in natural language, known as syntagmatic and paradigmatic associations, respectively. This is achieved by using a formal model of word meaning within the query expansion process. On ad hoc retrieval, our approach achieves statistically significant improvements in MAP (0.158) and P@20 (0.396) over our baseline model. The ERR@20 and nDCG@20 of our system was 0.249 and 0.192 respectively. Our results and discussion suggest that information about both syntagamtic and paradigmatic associations can assist with improving retrieval effectiveness on ad hoc retrieval.

## 1 Introduction

QUT's Quantum Interaction (QI) group conducts research into novel information retrieval methods that utilise richer representations of word meaning, and includes researchers from QUT and CSIRO. This paper details the methods used in our TREC 2012 Web track submission. Specifically, we focused on including information about word associations (i.e. term dependencies) to augment the query representation used within the ad hoc retrieval process.

### 1.1 Word Associations

Many existing retrieval approaches, including those using $tf.idf$ and those set within the unigram language modelling framework, do not explicitly take into account information about term dependencies. Term dependencies are those relationships between words that exist within natural language. Existing retrieval models that explicitly consider term dependencies mainly access information about first order relationships between words, known as syntagmatic associations. Syntagmatic associations are formed between words that co-occur near each other in text with a likelihood greater than chance. For example, in *A dog bit the mailman*, the term *dog* would have syntagmatic associations with *bit* and *mailman*, assuming they co-occur with *dog* above chance.

Within structural linguistics there is also another type of relationships between words, known as paradigmatic associations, that in combination with syntagmatic associations give words their meaning. The association between two words is paradigmatic if they can substitute for one another in a sentence without effecting the acceptability of the sentence. Typical examples are synonyms like *paper* - *article*, or related verbs like *eat* - *drink*. In the earlier example of *A dog bit the mailman*, the word *dog* could be switched for *snake*, demonstrating the paradigmatic association that exists between dog and snake.

## 2 Expanding Queries Using Word Associations

The following section outlines the approach we took to evaluate the efficacy of using information about syntagmatic and paradigmatic associations to augment query representations within the 2012 TREC Web track ad hoc retrieval task. A framework that allows information about both syntagmatic and paradigmatic associations to be explicitly modelled and combined exists within the *tensor encoding* (TE) model of word meaning [5].

The TE model automatically builds representations of words based on their co-occurrence patterns within a set of training documents. These representations are then used to model syntagmatic and paradigmatic associations. In our experiments, the set of training documents used to build the TE model representations are based on the set of $k$ top ranked pseudo relevant documents produced from a strong baseline model. Our baseline model is created using the following approach:

- The *ClueWeb09-Category B* documents are indexed using the 'indexing without spam' approach [7]. Each query is then issued to the Google retrieval service[1] and the top 60 retrieved documents are filtered using the spam filtered ClueWeb09-Category B index[2]. On average, 13 out of the 60 top ranked Google documents existed in this index. This filtered list is then padded, to create a list of 10,000 documents, using the list of documents returned from a search on the spam filtered index using a unigram language model. These rankings form our baseline submission (**QUTParaBline**). The use of Google as the search engine for the top ranked results and the filtering of spam web pages are likely to translate into a strong baseline. The process used to produce this baseline model is depicted in Figure 1.

The TE model has been used to underpin a formal query expansion technique, known as *tensor query expansion* (TQE) [6]. To augment a query representation within the TQE approach, an estimate of observing a vocabulary term $w$ given the query $Q$ is provided by the following conditional probability:

$$P(w|Q) = \frac{1}{Z}\left[\gamma S_{\text{par}}(Q,w) + (1-\gamma)S_{\text{syn}}(Q,w)\right], \tag{1}$$

where $w$ is any term in the TE vocabulary (formed from the set of $k$ pseudo relevant documents returned by our baseline model - QUTparaBline), $Q$ is the sequence of original query terms, $S_{\text{par}}(Q,w)$ is the measure of the strength of paradigmatic associations between $Q$ and $w$, $S_{\text{syn}}(Q,w)$ is the measure of the strength of syntagmatic associations between $Q$ and $w$, and

[1] `http://www.google.com`

[2] We had to limit the number of documents retrieved with Google to 60 because of Google's policies regarding the retrieval service at the time.

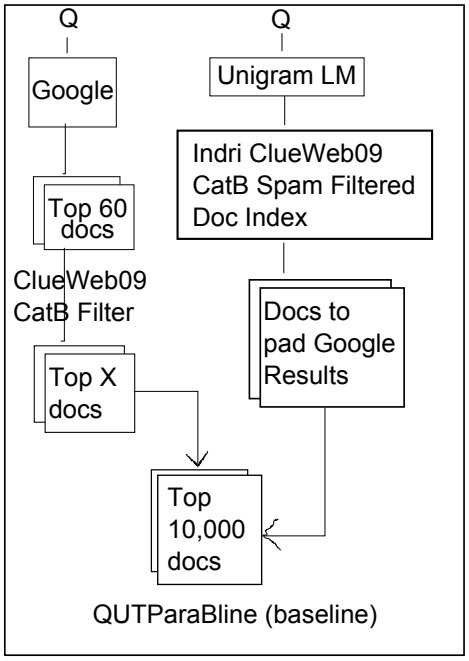


**Figure 1: Baseline: QUTParaBline.**

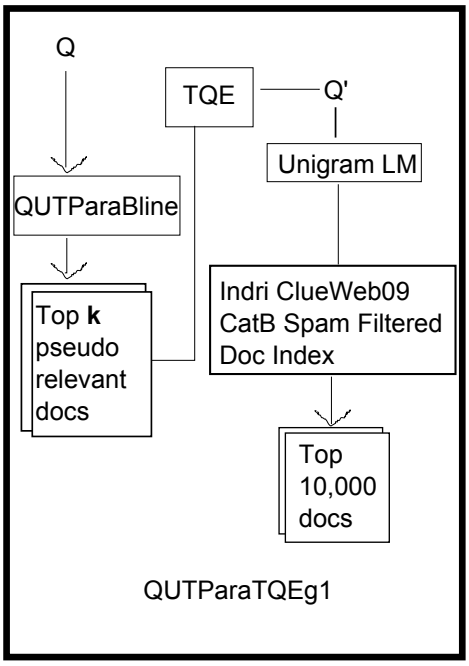


**Figure 2: QUTParaTQEg1.**

$\gamma \in [0, 1]$ mixes the paradigmatic $S_{\text{par}}()$ and syntagmatic $S_{\text{syn}}()$ measures, and $Z$ normalises the resulting distribution.

From these estimates an augmented query representation $Q'$ is created, as shown in Figure 2, and passed to a unigram language model to perform a final search on the spam filtered ClueWeb09 Category B index. The top ranked 10,000 documents are then submitted as part of our QUTParaTQEg1 run.

Tuning of the QUTParaTQEg1 system parameters, including $\gamma$ in Equation (1), was achieved by training on ERR@20 using the TREC web Track data sets from 2010 and 2011. The test parameter values used in the submission were *Number of feedback documents* equal to 19, *number of expansion terms* equal to 14, *original query weight* equal to 0.4 and *TE model mixing parameter* ($\gamma$) equal to 0.1.

# 3 Analysis of the runs

## 3.1 Setup

The document collection used in our experiments were based on a spam filtered version of the ClueWeb09-Category B corpus. Documents and queries were stopped using a standard INQUIRY stopword list [1] and stemmed using a Krovetz stemmer [3]. Queries submitted to Google were neither stopped nor stemmed.

Documents were indexed using the 'indexing without spam' method; the Waterloo spam list with threshold of 0.45 was used to estimate spam-likelihood of documents [2]. Indexing and retrieval approaches were implemented using the Indri toolkit [3]. The smoothing approach

[3]Available at `http://sourceforge.net/projects/lemur`

| | Graded Metrics | | Binary Metrics | |
|---|---|---|---|---|
| | ERR@20 | nDCG@20 | P@20 | MAP |
| unigramLM | 0.160 | 0.112 | 0.254 | .107 |
| MeanWT2012 | 0.187 | 0.123 | $0.284^{u}$ | - |
| QUTparaBline | $\mathbf{0.290}^{um}$ | $0.167^{um}$ | $0.305^{um}$ | $0.117^{u}$ |
| QUTparaTQEg1 | $0.249^{um}$ | $\mathbf{0.192}^{um}$ | $\mathbf{0.396}^{umb}$ | $\mathbf{0.158}^{ub}$ |
| | (-14.2%) | (+15%) | (+29.8%) | (+35%) |

**Table 1: Comparison of retrieval performance on TREC 2012 Web Track ad hoc retrieval task. The superscripts $u$, $m$, $b$ and $t$ indicate statistically significant differences (calculated using a paired t-test $p < 0.05$) over the unigram language model (unigramLM), the average performance of all TREC Web track participants (MeanWT2012), our baseline (QUTparaBline) and the TQE approach (QUTparaTQEg1), respectively. The best results for each evaluation measure appear in boldface. Brackets indicate the percentage change between QUTparaTQEg1 and QUTparaBline. Note that no value of MAP was provided for the average of all TREC 2012 Web Track submissions (MeanWT2012).**

and parameters used within the unigram language model were based on the Indri defaults.

### 3.2 Comparison of the Runs

In this section we compare the results of our two submissions to the TREC 2012 Web Track ad hoc task.

1. **QUTParaBline**: This run was produced by padding the ClueWeb09 Category B, spam filtered, top 60 Google results with the results returned by a unigram language model on the same spam filtered index (refer to Figure 1). This run forms our baseline.

2. **QUTParaTQEg1**: This run was produced by expanding the original TREC 2012 Web Track topics using TQE based on a set of $k$ pseudo-relevant documents produced by the baseline model (refer to Figure 2).

Table 1 compares the retrieval effectiveness of these runs (QUTparaBline, QUTparaTQEg1) along with the average effectiveness of all 48 TREC Web track submissions on the ClueWeb09 CategoryB collection (MeanWT2012), and a baseline unigram language model (unigramLM).

The results suggest that expanding query representations using TQE to model syntagmatic and paradigmatic associations of the original query terms can provide significant improvements over our strong baseline when binary metrics (i.e. MAP and P@20) are considered. No significant difference in retrieval effectiveness was noted on graded metrics (ERR@20 and nDCG@20).

The inability to achieve significant improvements on graded metrics may be related to the ability of the Google web search to return relevant documents high up the search order (i.e., in positions 1 and 2), when compared to the unigram language model's ability to achieve this. This can be seen by comparing graded metric scores of the unigram language (unigramLM) model and our baseline (QUTParaBline) in Table 1, and noting that the TQE model (QUTParaTQEg1) achieved significantly better P@20 than the QUTParaBline, which

indicates more relevant documents were returned in the top 20 by QUTParaTQEg1 than QUTParaBline.

The number of queries on which our submissions achieved a better than average performance when compared to other 2012 TREC Web track participants, for the 50 2012 TREC Web track topics, is shown in Table 2.

| | ERR@20 | nDCG@20 | P@20 |
|---|---|---|---|
| QUTparaBline | 34 | 30 | 33 |
| QUTparaTQEg1 | 29 | 31 | 36 |

**Table 2: Number of topics (out of 50) on which our baseline (QUTParaBline) and TQE approach (QUTParaTQEg1) outperformed the average of all 2012 TREC Web track participants, for the ERR@20, nDCG@20 and P@20 measures.**

## 3.3 Further Analysis

### 3.3.1 Comparison with Other teams

The average, worst and best performance using ERR@20, nDCG@20 and P@20 across all 48 submissions to the 2012 TREC Web track is provided by the NIST organisers. The average is shown as MeanWT2012 in Table 1; both of our submissions (QUTparaTQEg1 and QUTparaBline) produce statistically significant improvements (using a paired t-test and 95% confidence interval) in all metrics when compared to MeanWT2012.

### 3.3.2 Success and Failures

From comparing the query-by-query retrieval performance of the QUTparaTQEg1 system with the best and worst performance across all systems, we can see that our paradigmatically enhanced retrieval approach does well on the following queries, in which it achieves the best (or equal best) ERR@20 or nDCG@20 when compared to all other TREC Web track competitors.

1. Topic 184: Civil right movement
2. Topic 192: Condos in Florida
3. Topic 197: Idaho state flower

However, the queries where the worst non-zero performance (ERR@20 or nDCG@20) was experienced included:

1. Topic 172: becoming a parallegal

The queries where both nDCG@20 and ERR@20 were zero for the QUTParaTQEg1 system (and the best performance across all participants was non-zero), included:

1. Topic 157: The beatles rock band
2. Topic 162: dnr
3. Topic 163: arkansas

4. Topic 167: barbados
5. Topic 170: scooters
6. Topic 179: black history
7. Topic 188: internet phone service
8. Topic 189: gs pay rate

### 3.4 QUTPara on Diversity

Approaches to result diversification are commonly distinguished between those that explicitly diversify document rankings trying to estimate queries and documents intents, and those that implicitly diversify the rankings, without explicitly predicting query intents and the coverage of documents with respect to those intents.

While we did not make any design choices explicitly aimed at improving the diversity of results, we hypothesis that the TQE approach (QUTparaTQEg1) may to some extent, diversify document rankings as it considers different forms of word associations to build query representations (i.e. syntagmatic and paradigmatic associations). It is these two different sources of word associations that provide, in our hypothesis, implicit diversification of results. In addition, our baseline (QUTparaBline) may also provide search result diversity in its top ranking as it is likely that Google caters for diversification.

Results obtained by our submissions when evaluated on the diversity task are shown in Table 3. These results show that both our submissions perform better than the average. While TQE (QUTParaTQEg1) provides search results that exhibit statistically significant improvements in P-IA@20 over our baseline (QUTparaBline), our baseline obtains the highest scores for $\alpha$-nDCG@20 and ERR-IA@20 among our submissions: the reason for such variability is yet unclear, but may be rooted in the parametrization of $\alpha$-nDCG@20 and ERR-IA@20 [4].

| | $\alpha$-nDCG@20 | P-IA@20 | ERR-IA@20 |
|---|---|---|---|
| MeanWT2012 | 0.476 | 0.213 | 0.364 |
| QUTparaBline | $\mathbf{0.527}^{m}$ | 0.226 | $\mathbf{0.419}^{m}$ |
| QUTparaTQEg1 | 0.498 | $\mathbf{0.286}^{mb}$ | 0.382 |

**Table 3: Comparison of performance on TREC 2012 Web Track diversity task. The superscripts $m$, $b$ and $t$ indicate statistically significant differences (calculated using a paired t-test, $p < 0.05$) over the average performance of all TREC Web track participants (MeanWT2012), our baseline (QUTparaBline) and TQE approach (QUTparaTQEg1), respectively. The best results for each evaluation measure appear in boldface.**

## 4 Conclusions

Many existing retrieval approaches do not explicitly take into account information about word associations, commonly known as term dependencies in information retrieval. Those that do, primarily model dependencies known as syntagmatic associations. Our approach explicitly

models both syntagmatic and paradigmatic associations within the retrieval process. These associations have been argued by structural linguists to form the meaning of words.

In our submissions for TREC 2012 Web Track, we used a formal model of word meaning to enhance the query representation of the topic titles. This approach, known as *tensor query expansion* (TQE) was designed to investigate the benefits of including information about syntagmatic and paradigmatic associations within the retrieval process. Our experiments on ad hoc retrieval using the 2012 TREC Web Track topics, provide evidence to suggest that including both syntagmatic and paradigmatic information can significantly improve MAP and P@20 when compared to a strong baseline. Our TQE approach and baseline submission both achieved significant improvements in retrieval effectiveness when compared to the average scores of all 2012 TREC Web track submissions[4].

Even though the overall performance of our approach was significantly better than average, a query by query analysis suggests that not all of the queries are effectively expanded. Continued investigation into the use of the TQE approach may provide further insights into the role word associations can play in providing effective query representations for use in document retrieval.

## 5 Acknowledgements

We would like to thank the QUT HPC team for the use of the computational resources to perform our experiments and Dr. Paul Thomas for feedback on the use of the Google retrieval service.

## References

[1] J. Allan, M. E. Connell, W. B. Croft, F. F. Feng, D. Fisher, and X. Li. Inquery and trec-9. In *Proceedings of TREC 2000*, TREC 2000, 2000.

[2] G. V. Cormack, M. D. Smucker, and C. L. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Inf. Retr.*, 14(5):441–465, Oct. 2011.

[3] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 191–202, New York, NY, USA, 1993. ACM.

[4] T. Leelanupab, G. Zuccon, and J. M. Jose. A comprehensive analysis of parameter settings for novelty-biased cumulative gain. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, 2012.

[5] M. Symonds, P. Bruza, L. Sitbon, and I. Turner. Modelling word meaning using efficient tensor representations. In *PACLIC 25*, pages 313–322, 2011.

[6] M. Symonds, P. Bruza, L. Sitbon, and I. Turner. Tensor query expansion: a cognitive based relevance model. In *Proceedings of the 16th Australasian Document and Computing Syposium (ADCS 2011)*, pages 87–94. RMIT University(Melbourne), 2011.

[4] No average MAP was provided when the results were initially released.

[7] G. Zuccon, A. Nguyen, T. Leelanupab, and L. Azzopardi. Indexing without spam. In *Proceedings of the 16th Australasian Document and Computing Syposium (ADCS 2011)*, 2011.