University of Glasgow at the Web Track: Dynamic Application of Hyperlink Analysis using the Query Scope

Vassilis Plachouras[1], Fidel Cacheda[2], Iadh Ounis[1], and
Cornelis Joost van Rijsbergen[1]

[1]University of Glasgow, G12 8QQ Glasgow, UK
[2]University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

Abstract

This year, our participation to the Web track aims at combining dynamically evidence from both content and hyperlink analysis. To this end, we introduce a decision mechanism based on the so-called *query scope* concept. For the topic distillation task, we find that the use of anchor text increases precision significantly over content-only retrieval, a result that contrasts with our TREC11 findings. Using the query scope, we show that a selective application of hyperlink analysis, or URL-based scores, is effective for the more generic queries, improving the overall precision. In fact, our most effective runs use the decision mechanism and outperform significantly the content and anchor text retrieval. For the known item task, we employ the query scope in order to distinguish the named page queries from the home page queries, obtaining results close to the content and anchor text baseline.

1 Introduction

Our participation in both tasks of the Web track is a continuation of our TREC11 evaluation of a modular probabilistic framework for Web Information Retrieval, which integrates content and link analysis [6]. This year, we introduce a new notion called *query scope*, which is employed in order to decide dynamically about the most appropriate combination of content and hyperlink analyses. Our assumption is that a query-based application of link analysis performs better than a uniform approach, where the same method is applied indifferently for all queries.

The estimation of the query scope is based on statistical evidence obtained from the set of retrieved documents. The query scope is used as an input into a simple decision mechanism, with which we select the most appropriate retrieval approach for each query. The approaches we employ are content-only retrieval, retrieval based on the content of documents and the anchor text of their incoming links, and a combination of the last approach with a score obtained from the URL length of a document, or from the Static Utility Absorbing Model, a hyperlink analysis model [7].

This year, we find that the combination of content and anchor text retrieval is highly effective, increasing the precision over content-only retrieval. This contrasts with our findings from last year's TREC11 topic distillation task, where the content-only retrieval outperformed any other approach for the topic distillation task. In addition, we find that using a URL length-based score for the most generic queries increases precision for the same task.

The known item task for this year is a mixture of named and home page queries. The query scope is used to detect the probable type of each query, and apply an appropriate retrieval approach. For example, we assume that home page queries could benefit from the combination of content and anchor text with the URL length score. We find that content and anchor text is still a very effective retrieval approach for this task.

The rest of the paper is organised as follows. In Section 2, the query scope is defined. Sections 3 and 4 contain a description of our official runs and a set of additional runs for the topic distillation task. In Section 5, we present our experiments for the known item task, while Section 6 contains our conclusions from this year's participation in the Web track.

2 The Query Scope

Assuming that not all queries benefit from the same retrieval approach, we need to find which of the available approaches is most appropriate for a specific query. Hence, we introduce a decision mechanism that will associate to each query the most appropriate approach. For example, for specific queries we employ a content-only retrieval, while for more generic queries, we use evidence from the hyperlinks, or the URLs. The decision mechanism is based on a composite measure, the query scope, which addresses three important statistical aspects of the set of retrieved documents.

The first aspect addressed is related to the number of retrieved documents. We assume that for the more generic queries, there will be many documents that contain all the query terms. In these cases, the queries address a topic that is widely covered in the collection. Therefore, evidence from hyperlink analysis may be more useful in detecting high quality documents, or homepages of relevant sites.

The $query\_extent$ is the number of retrieved documents that contain all the query terms, normalised between 0 and 1 by dividing with a given fraction α of the total number of documents in the test collection:

$$query\_extent = \min\left(\frac{\{\text{number of retrieved documents containing all query terms}\}}{\alpha}, 1\right) \tag{1}$$

The normalisation is introduced as most of the queries tend to retrieve only a small fraction of documents from the collection, and therefore dividing by α leads to a better distributed measure. In our experiments, we normalised the query extent by dividing with 1% of the number of documents in the collection.

The second aspect is about finding whether there are sites devoted to the query's topic. If there are such sites, we expect that they will contain a high number of documents with all the query terms, and that their homepages will be more useful than other documents. Similarly to the approaches used in [5, 8], we assume that a site is defined by the domain of the document's URL, and we group the documents according to their domain.

We denote by $size_j$ the number of documents from the jth site. In addition, let μ_{size} and σ_{size} be the average and the standard deviation respectively of $size_j$ for $j \in [1, n]$, where n is the number of the retrieved sites. We define the $result\_extent$ as the number of sites for which the size is higher than $\mu_{size} + 2 \times \sigma_{size}$:

$$result\_extent = \{\text{number of sites for which } size_j > \mu_{size} + 2 \times \sigma_{size}\} \tag{2}$$

The third aspect is related to the distribution of *root*, *subroot* and *path* documents among the top ranks. Kraaij et al. [4] have classified the documents into four types according to the type of their URL:

- *root* documents (e.g. http://www.dcs.gla.ac.uk/)
- *subroot* documents (e.g. http://www.dcs.gla.ac.uk/ir/)
- *path* documents (e.g. http://www.dcs.gla.ac.uk/ir/projects/)
- *file* documents (e.g. http://www.dcs.gla.ac.uk/ir/people.html)

We expect that for the home page finding queries, there will be a higher number of documents of the first three types distributed in the top ranks, than for named page finding or topic distillation queries. For this reason, we employ the sum of the reciprocal ranks of documents from the first three categories as an indication of the user's intent to retrieve a home page, so that both the ranks and the number of these documents are taken into account.

In order to obtain the sum of reciprocal ranks, we rank the documents according to their content, and we denote the ith document by d_i. Then, if d_i is either a root, a subroot, or a path, we set $\text{RR}(i)$ to the reciprocal rank of d_i, otherwise we set it equal to zero:

$$\text{RR}(i) = \begin{cases} \frac{1}{i} & \text{if } d_i \text{ is a } \textit{root, subroot}, \text{ or } \textit{path} \text{ document} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The sum of the reciprocal ranks is given by:

$$rank\_sum = \sum_{i=1}^{k} \text{RR}(i) \tag{4}$$

In all our experiments we used $k = 100$.

Having defined the three components of query scope in Equations (1), (2) and (4) respectively, we can proceed with defining a decision mechanism that will employ these measures. Depending on the Web track task, we use the most appropriate of the components defined above, in order to decide how to combine the content and the hyperlink analysis. For the topic distillation task, we employ the $query\_extent$ and the $result\_extent$, while for the known item finding task, we use the $query\_extent$ and the $rank\_sum$.

More specifically, for the topic distillation task, the decision mechanism is shown in Table 1. The queries that have high $query\_extent$ and $result\_extent$ (case I), are treated as more generic. On the contrary, the queries, which result in low values for the query scope components (case IV), are considered to be specific. For the rest of the queries (cases II and III), we cannot say with confidence whether they are generic or specific queries. Both thresholds t_{qe} and t_{re} are determined experimentally.

	$result\_extent \geq t_{re}$	$result\_extent < t_{re}$
$query\_extent \geq t_{qe}$	case I (generic queries)	case III (unknown)
$query\_extent < t_{qe}$	case II (unknown)	case IV (specific queries)

Table 1: The decision mechanism for the topic distillation task

For the known item finding task, our aim is to distinguish between the named page finding queries and the home page finding queries. We expect that the home page finding queries will retrieve more documents containing all the query terms and that there will be more root, subroot or path documents distributed in the top ranks. Therefore, in the decision mechanism for the known item finding task, we employ the $query\_extent$ and the $rank\_sum$, as shown in Table 2. The named page queries are expected to have low values for both $rank\_sum$ and $query\_extent$ (case IV), while home page queries should have either high $rank\_sum$, or high $query\_extent$ (cases I, II and III). Again, both thresholds t_{qe} and t_{rs} are determined experimentally.

	$rank\_sum \geq t_{rs}$	$rank\_sum < t_{rs}$
$query\_extent \geq t_{qe}$	case I (home page query)	case III (home page query)
$query\_extent < t_{qe}$	case II (home page query)	case IV (named page query)

Table 2: The decision mechanism for the known item finding task

3 Topic Distillation Task

Our aim is to investigate the appropriateness of the query scope and the corresponding decision mechanism described in Section 2. We test four different retrieval approaches: content-only, content and anchor text, content and anchor text with URL length, content and anchor text with the Static Utility Absorbing Model [7].

In all runs, we removed stop words and applied stemming during indexing. For the content analysis, we used the weighting model $PL2$ from Amati and van Rijsbergen's Divergence From Randomness (DFR) framework [1], with $c = 1.28$. As opposed to last year, the estimation of the c value is done using a new parameter-tuning methodology for term frequency normalisation, which is collection-independent [3].

The first run, uogtd1c, is a content-only baseline, where only the content of the documents is used. For the second run, uogtd2ca, we test the usefulness of combining the content of documents with the anchor text of their incoming links. Comparing the results of the first two runs, we can see that there is a significant improvement from using anchor text, for both precision at 10 and R precision (see Table 3).

With the third run, we introduce a simplified version of the decision mechanism, as defined for the topic distillation task in Section 2. For the queries where $result\_extent < t_{re}$, we use the content of documents only. For the rest of the queries, where $result\_extent \geq t_{re}$, we employ the content and anchor text of documents.

Official run	Prec. at 10	R precision	Features
uogtd1c	0.0680	0.0730	content-only
uogtd2ca	0.1020	0.1325	content and anchor text
uogtd3cas	0.0820	0.1053	using dynamically content and anchor text
uogtd4cahs	**0.1140**	**0.1391**	using dynamically content, anchor text and URL length
uogtd5cass	0.1080	0.1361	using dynamically content, anchor text and SUAM

Table 3: Topic distillation official results

Therefore, in this configuration, we only use $result\_extent$, for which the threshold is experimentally set to $t_{re} = 7$ (Table 4). In Table 5, we show the number of queries for which each retrieval approach was applied, along with the number of queries for which the applied approach was at least as effective as the alternative retrieval approach.

$result\_extent \geq 7$	$result\_extent < 7$
content and anchor text	content-only

Table 4: The simplified decision mechanism for run uogtd3cas

Approach	Applied for # queries	Effective for # queries (Pr. at 10)	Effective for # queries (R pr.)
content-only	19	10	12
content and anchor text	31	28	29

Table 5: The number of queries for which each approach was applied in run uogtd3cas, and the number of queries for which the applied approach was at least as effective as the alternative approach.

Comparing the effectiveness of this approach (i.e. uogtd3cas) with the first two runs, we can see that both precision at 10 and R precision are between the corresponding measures of the content-only and the content and anchor text runs. However, it is not clear whether this is due to a failure of the decision mechanism, or due to the poor effectiveness of the content-only retrieval.

In the fourth run, uogtd4cahs, we employ a more fine-grained classification by using both $query\_extent$ and $result\_extent$. For the most generic queries (case I), we use both content and anchor text, and we re-rank the top 1000 documents by taking into account their URL length, in order to boost the homepages. Let sc_i be the content analysis score of the ith document. In addition, let $urlpath\_len_i$ be the length in characters of the ith document's URL path[1]. The final score for this document will be:

$$score_i = sc_i \times \frac{1}{\log_2(urlpath\_len_i + 1)} \tag{5}$$

The decision mechanism used for this run is shown in Table 6. The thresholds t_{re} and t_{qe} were set experimentally to the values 7 and 0.5 respectively. Table 7 contains the number of queries for which each retrieval approach was applied, along with the number of queries for which the applied approach was at least as effective as the others.

	$result\_extent \geq 7$	$result\_extent < 7$
$query\_extent \geq 0.5$	content, anchor text and URL length	content-only
$query\_extent < 0.5$	content and anchor text	content-only

Table 6: The decision mechanism for run uogtd4cahs

Approach	Applied for # queries	Effective for # queries (Pr. at 10)	Effective for # queries (R pr.)
content-only	19	10	12
content and anchor text	16	15	15
content, anchor text and URL length	15	14	14

Table 7: The number of queries for which each approach was applied in run uogtd4cahs, and the number of queries for which the applied approach was at least as effective as the others.

As shown in Table 3, the above approach improves both precision at 10 and R precision significantly over run uogtd3cas. Moreover, looking in Table 7, we find that both combinations of content with anchor text, and content with anchor text and URL length are very effective for the queries on which they were applied respectively. On the other hand, content-only retrieval is not very appropriate for the selected queries.

For our last official run of the topic distillation task, uogtd5cass, we combine the content and anchor text of documents with the Static Utility Absorbing Model (SUAM) [7], a hyperlink analysis approach. Let sc_i be the

[1] For example, for the URL http://trec.nist.gov/data/intro_eng.html, the path is data/intro_eng.html

content analysis score and sc_{AMi} be the Absorbing Model score for the ith document. The final score is given by:

$$score_i = sc_i^A \times (-\log_2(sc_{AMi}))^B \tag{6}$$

where $A = B = 1$. The Static Utility Absorbing Model scores documents according to both their incoming and outgoing links, aiming to boost the key entry points for a given topic. The decision mechanism employed is shown in Table 8, where the thresholds t_{re} and t_{qe} are set to 7 and 0.5 respectively. In all cases, we employ the content and anchor text of documents, while SUAM is only used for the most generic queries.

When comparing this run to uogtd2ca, where content and anchor text is used for all the cases, we note that precision at 10 increases slightly (Table 3), due to improvement in 2 out of the 15 most generic queries, for which SUAM is applied (Table 8). For the rest of the queries, effectiveness remains stable. If we consider R precision, we find that SUAM results in improvements for 2 queries, but is also detrimental for 3 queries on which it is applied.

	$result\_extent \geq 7$	$result\_extent < 7$
$query\_extent \geq 0.5$	content, anchor text and SUAM	content and anchor text
$query\_extent < 0.5$	content and anchor text	content and anchor text

Table 8: The decision mechanism for run uogtd5cass

4 Additional Topic Distillation Experiments

After obtaining the official results and the relevance assessments, we run additional experiments for the topic distillation task. The aim is to learn more about the effectiveness of the query scope and the decision mechanism.

In the additional runs, we test separately the effectiveness of $query\_extent$ and $result\_extent$. To facilitate the analysis, employing the three retrieval approaches used in uogtd4cahs (Table 7), we use the two most effective ones, in terms of average precision at 10 over all queries. As shown in Table 9, these are content and anchor text retrieval, and content with anchor text and the URL length-based score (Equation (5)).

Employing the decision mechanism as shown in Table 1, we use Equation (5) for the cases I and II and only the content and anchor text for the cases III and IV. This corresponds to employing only the $result\_extent$, for which the threshold t_{re} takes the values $4, 7, 10, 13$ (runs re$i$ in Table 9). On the other hand, if we use Equation (5) for the cases I and III, and we employ only the content and anchor text of documents for the cases II and IV, then this approach corresponds to applying only the $query\_extent$, for which the threshold t_{qe} takes the values $0.25, 0.45, 0.50, 0.55, 0.65$ (runs qe$i$ in Table 9).

Run	Threshold	Prec. at 10	R precision
content-only (uogtd1c)		0.0680	0.0730
content and anchor text (uogtd2ca)		0.1020	0.1325
content, anchor text and URL length		0.1400	0.1369
re1	$t_{re} = 4$	0.1360	0.1340
re2	$t_{re} = 7$	0.1440	0.1428
re3	$t_{re} = 10$	**0.1480**	0.1528
re4	$t_{re} = 13$	0.1240	0.1395
qe1	$t_{qe} = 0.25$	0.1280	0.1547
qe2	$t_{qe} = 0.45$	0.1340	**0.1657**
qe3	$t_{qe} = 0.50$	0.1340	**0.1657**
qe4	$t_{qe} = 0.55$	0.1300	0.1635
qe5	$t_{qe} = 0.65$	0.1260	0.1542

Table 9: Additional results for the topic distillation

From the additional results, we can see that by employing $result\_extent$, we achieve higher precision at 10 for two out of the four threshold values tested. More specifically, for $t_{re} = 10$, we obtain 0.1480 precision at 10. When we use $query\_extent$, precision at 10 is between the average precision at 10 of the two most effective retrieval approaches employed. Now, if we consider R precision, both $result\_extent$ and $query\_extent$ lead to improvements over the best performing retrieval approach. In addition, $query\_extent$ is more effective than $result\_extent$, for the tested threshold values. When t_{qe} is equal to 0.45 or 0.50, we obtain 0.1657 R precision.

5 Known Item Finding Task

For the known item finding task, we employ the query scope in order to discriminate between home page and named page queries. We assume that the most effective method for named page queries is to employ content and anchor text retrieval. For home page queries, we use content and anchor text retrieval, and re-rank all the retrieved documents by applying the formula from Equation (5).

Official run	Aver. Recip. Rank.	Found in top 10	Not Found	Features
uogki1c	0.363	167 (55.7%)	82 (27.3%)	content-only
uogki2ca	**0.615**	238 (79.3%)	34 (11.3%)	content and anchor text
uogki3cah	0.273	117 (39.0%)	92 (30.7%)	content and anchor text and URL length
uogki4cahs	0.595	227 (75.7%)	30 (10.0%)	using dynamically content, anchor text and URL length

Table 10: Known item finding official results

In the first run, uogki1c, we use only the content-only retrieval. The used weighting model is $PL2$, as in the case of topic distillation, but with $c = 1$ this time. As expected from last year's named page finding task [2], the content-based approach is not the most efficient approach (see Table 10). In the second run, uogki2ca, where the document's content and the anchor text of the incoming links are used, the effectiveness increases significantly.

For the third run, uogki3cah, we also employ the URL length, in the same way as in Equation (5), in order to boost the home pages, which tend to have shorter URLs. Although this approach is not appropriate for all queries, it is expected to be effective at least for the home page finding queries.

Before proceeding with the fourth run, we will look at the effectiveness of content and anchor text retrieval, and content with anchor text and URL length, across the sets of named page and home page finding queries. For the named page queries, content and anchor text retrieval results in 0.613 average reciprocal rank, and for the home page queries the average reciprocal rank is 0.617. On the other hand, content with anchor text and URL length is not so stable across the two sets of queries. For the named page queries, the average reciprocal rank is 0.060, while for the home page finding task it is 0.487. These results show that content and anchor text retrieval is a very robust approach for the known item finding task, and indicate that there is no significant gain in trying to combine different methods.

However, if we manually select the best approach for each query, then we get 0.680 average reciprocal rank across the two sets of queries. For the named page finding queries, we would get 0.613 and for the home page finding queries we would get 0.747. These figures show that there is room for improvement if a successful combination is found.

Our last official run intends to simulate automatically the above manual selection. Therefore, we employ the query scope as defined in Table 2 for the known item task. We use two components of the query scope, the $query\_extent$ and the $rank\_sum$, with threshold values set experimentally to $t_{rs} = 1$ and $t_{qe} = 0.8$ as shown in Table 11. For all queries, the content and the anchor text of the incoming links is used, and for the most generic ones, where $t_{rs} \geq 1$ and $t_{qe} \geq 0.8$, the URL's length is used as defined in Equation (5). As shown in Table 10, precision is close to that of run uogki2ca, where content and anchor text is used.

We find that the average reciprocal rank of named page queries is 0.554, while for the home page queries it is 0.636. More specifically, the URL length is employed for 56 queries, of which 19 were named page queries and 37 were home page queries. For those 19 named page queries where the URL length is employed, there is no additional improvement. However, for 11 out of the 37 home page queries where the URL length is used, there is an improvement over using only the anchor text. These results show that our decision mechanism succeeds in increasing the effectiveness for the home page queries, but on the other hand, it does not benefit named page queries. More work is needed in order to refine the decision mechanism and improve its effectiveness, compared to that of the manual selection of the most appropriate approaches per query.

	$rank\_sum \geq 1$	$rank\_sum < 1$
$query\_extent \geq 0.8$	content, anchor text and URL length	content, anchor text and URL length
$query\_extent < 0.8$	content, anchor text and URL length	content and anchor text

Table 11: The decision mechanism for run uogki4cahs

6 Conclusions

We introduce a dynamic approach for combining content and hyperlink analysis, based on the query scope, a composite measure for quantifying three aspects of how appropriate a query is for hyperlink analysis.

For the topic distillation, we find that employing the anchor text of incoming links outperforms significantly a content-only retrieval. This result contrasts with our findings from TREC11, where the content-only retrieval was the most effective approach for topic distillation. In addition, URL information is very effective in identifying the relevant homepages. As for the dynamic combination of different retrieval approaches, it appears that the decision mechanism is a very effective method, leading to significant improvements in our official results. Moreover, when the most effective individual approaches are employed, we find that the $result\_extent$ is more effective with respect to precision at 10, while $query\_extent$ achieves better results when R precision is used for the evaluation.

For the known item finding task, the content and anchor text retrieval is a robust approach, independently of the query type. However, the URL information is useful for the home page queries, and if it is applied appropriately, it results into significant improvement. Our decision mechanism performed nearly as well as the content and anchor text baseline. It resulted in increased effectiveness among the home page queries, while it didn't benefit the named page queries.

In conclusion, we have shown that by employing simple statistical mechanisms, it is possible to improve the retrieval effectiveness by combining dynamically evidence from content and hyperlink analysis.

Acknowledgements

The work of the first, third and fourth authors is funded by a UK Engineering and Physical Sciences Research Council (EPSRC) project grant, number GR/R90543/01. The project funds the development of the Terrier Information Retrieval framework (url: http://ir.dcs.gla.ac.uk/terrier).

The work of the second author has been partially supported by the 'Comisión Interministerial de Ciencia y Tecnologia' (CICYT) of the Spanish government, under project TIC2001-0547 and by the 'Fundación Caixa Galicia' (Beca curso 2002/2003 para Estudios de Postgrado en Universidades y en Centros de Investigación de Excelencia Académica).

References

[1] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 40(4):1–33, 2002.

[2] N. Craswell and D. Hawking. Overview of the TREC-2002 Web Track. In *NIST Special Publication: 500-251 The 11th Text REtrieval Conference (TREC 2002)*, pages 86–93. NIST, 2002.

[3] B. He and I. Ounis. A study of parameter tuning for term frequency normalization. In *Proceedings of the 12th international Conference on Information and Knowledge Management (CIKM)*, pages 10–16. ACM Press, 2003.

[4] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, 2002.

[5] K. L. Kwok, P. Deng, N. Dinstl, and M. Chan. TREC2002 Web, Novelty and Filtering Track Experiments using PIRCS. In *NIST Special Publication: SP 500-251 The 11th Text Retrieval Conference (TREC)*, pages 520–528, NIST, 2002.

[6] V. Plachouras, I. Ounis, G. Amati, and C. J. van Rijsbergen. University of Glasgow at the Web track of TREC 2002. In *NIST Special Publication: SP 500-251 The 11th Text Retrieval Conference (TREC)*, pages 645–651, NIST, 2002.

[7] V. Plachouras, I. Ounis, and G. Amati. A Utility-oriented Hyperlink Analysis Model for the Web. In *Proceedings of the 1st Latin Web Conference*, pages 123–131. IEEE Press, 2003.

[8] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, and L. Zhao. THU TREC 2002: Web Track Experiments. In *NIST Special Publication: SP 500-251 The 11th Text Retrieval Conference (TREC)*, pages 586–594, NIST, 2002.