# University of Glasgow at the Web track of TREC 2002

Vassilis Plachouras, Iadh Ounis, Gianni Amati*, and C.J. Van Rijsbergen

Department of Computing Science
University of Glasgow
Glasgow G12 8QQ
{vassilis, ounis, gianni, keith}@dcs.gla.ac.uk

**Abstract**

The aim of our participation in the topic distillation and the named page finding tasks of the Web track is the evaluation of a well-founded modular probabilistic framework for Web Information Retrieval, which integrates content and link analyses. The link analysis component of the framework employs a new probabilistic approach, called the Absorbing Model, for calculating a measure of popularity for documents induced from the Web graph.

## 1 Introduction

Both topic distillation and named page finding tasks require a high precision in the ranking, since there are few relevant documents for the topic distillation task, which reduce to one for the named page finding task. Results of TREC last year [5] have shown that the link structure upon the Web hardly combines with the content analysis, at least for the topic relevance task. Similarly, the length of the URL and other heuristics have shown to be more effective than the explicit use of the link structure for the homepage finding task. In general, it has been found that pure link structure analysis does not enhance retrieval effectiveness.

Our participation in the Web track focuses on a novel way to integrate content with link analysis. We propose a method for combining content and links in a sound and non-parametric way, as well as a dynamic spreading activation mechanism to be applied on top of this methodology. We evaluate our proposals by extending the framework used in TREC10 [1] in a modular way. The refined content analysis module is integrated with a new link analysis module. Moreover, the system is extended with a new query expansion module, which can be enabled, or disabled during the retrieval process.

The preliminary results reported in this paper show that link analysis may benefit the ranking, enhancing the overall retrieval effectiveness. However, its success strongly depends on the basic probabilistic link model used, as well as on the methodology employed to build the link structure graph. Our approach differs from that used in PageRank [4] and it can be seen as a dynamic modification of the link structure according to the content and thus, indirectly, to the query. Moreover, results show that different strategies are required for optimal topic distillation and named page finding effectiveness.

The rest of the paper is organised as follows. In Section 2, the proposed approaches and their rationale are described. Section 3 contains a detailed description of the runs submitted, and in Section 4 we present a preliminary analysis of the results.

* Also affiliated to Fondazione Ugo Bordoni, Rome.

# 2 Integration of content and link analyses

We introduce a new probabilistic model, which we call the Absorbing Model, for analysing the link structure of Web documents [2]. This method provides us with a measure of authority for documents, based on the probability distribution of accessing the states of a Markov chain, which is induced from the Web graph. This probability distribution corresponds to the principal eigenvector of the adjacency matrix of the Web graph, and it is computed by an iterative but necessary *converging* process. We apply this method either statically or dynamically.

## 2.1 A parameter-free link model: the Absorbing model

PageRank [4] connects any couple of documents in the collection through a virtual link, with a very small transition probability, which is a parameter for the algorithm. In the terminology of Markov chains, all documents on the Web become ergodic states, which are treated as nodes belonging to a unique large cluster. This cluster guarantees that all documents receive a probability different from zero (the authoritative score), which is the probability that the node is reached by any possible random walk through the nodes of the collection. Our Markov chain model offers a completely different solution. We transform the original graph to make all states ergodic and have many different clusters instead of one. The transition probabilities are then computed accordingly. In this transformation process, we do not need the use of parameters to obtain the final authoritative scores.

## 2.2 Static application of the Absorbing model

For the static application of the Absorbing Model (SAM), we first calculate an absolute authority score for each document in the collection during indexing. After retrieval is performed, this authority score is combined with the content based score in a parametric way. This approach is similar to the PageRank philosophy.

## 2.3 Dynamic application of the absorbing model

The dynamic application of the Absorbing model, which is called Dynamic Absorbing Model (DAM), seems to be promising. It is applied only on the top retrieved documents and aims to provide a dynamic authority measure. The retrieved documents are returned by the application of a classical content analysis model. Using a high quality content retrieval module, we assume that most of the authoritative documents will be among the top ranked documents, and thus the application of link analysis may change, but not drastically, the ranking of documents. Following this idea, the DAM is applied on the set $B$ of the top $|B|$ ranked documents. The priors are initialised using the scores of the content analysis. Among the top retrieved documents, we select the subset $A \subset B$, with $0 \leq |A| < |B|$, with the highest content scores. We suppose that the documents in this set $A$ should be more authoritative than the remaining ones in $B - A$. Therefore, we modify the subgraph of the subset $A$ of the top ranked documents of $B$ by removing their outgoing links. In this way, the documents in the set $B - A$ that might be relevant are boosted according to the link structure of $|B|$, while documents in $A$ do not loose their authority score but they may inherit some more from $B - A$. The application of DAM does not involve any training, which makes it suitable for large collections of documents, such as the Web.

## 2.4 Spreading Activation

In addition to the Absorbing Model, we propose a dynamic spreading activation mechanism for finding the best entry points for topics on the Web.

Spreading activation is a well known mechanism used in hypertext and web retrieval systems [8, 6], where a fraction of the Retrieval Status Value (*RSV*) of each document propagates to the documents linked by it, assuming that documents linked by other relevant documents, are possibly relevant as well. We adapt this mechanism in order to fit our need for identifying the best entry points for a topic. In our method, which we call Static Spreading Activation (SSA), the propagation of *RSV*s is performed only when the documents linked belong to the same domain and the source document of the link is deeper in the document tree than the destination document.

### 2.5 Query-biased Spreading Activation

The above spreading activation mechanism is refined by allowing the fraction of the $RSV$ that propagates to vary, depending on a measure of the query specificity. This measure, which we call *query scope*, is a hybrid probabilistic measure that depends on both collection statistics and the external conceptual structure provided by WordNet [7]. It is based on the assumption that a generic query consists of terms that correspond to generic concepts in a conceptual hierarchy, and also that occur frequently in the collection. In the dynamic version of the spreading activation mechanism, which we call Dynamic Spreading Activation (DSA), the query scope is used to adjust the effect of the *RSV* propagation, according to the assumption that generic queries may benefit more from the link structure analysis, and therefore from the propagation of a larger fraction of the *RSV*s between linked documents. During indexing, each term appearing in the collection is mapped to one or more concepts in WordNet and a probabilistic measure of the specificity of the term is calculated. During retrieval, the sum of the query term specificity measures, normalised by the length of the query, represents a probabilistic measure for the specificity of the query, which is used to refine the application of the spreading activation mechanism.

## 3 Description of experiments

### 3.1 Indexing

For indexing the collection, a standard stop word list is used and Porter's stemming algorithm is applied. Moreover, in all runs, except for `uog05tad` and `uog06c`, the documents are modified by doubling the occurrences of terms in titles and by adding to each document the anchor text of its incoming links. For the run `uog06c`, the documents are not modified, while for the run `uog05tad`, the inverted file is augmented with additional information on whether a term appears in the anchor text of a document or in its body. For all runs, the weighting formulas used for calculating the *RSV*s are a *variant* of $I(n_e)B2$, which we call $I(n_e)C2$, except for runs `uog04cta2dqh` and `uog09cta2`, where the formula $I(n_e)B2$ is used [3].

### 3.2 Topic Distillation

The static Absorbing Model (SAM) is applied only in the run `uog01ctaialh` for the topic distillation task. In more details, the probability distribution of accessing the states of a modified Markov chain obtained from the graph of the collection is calculated during indexing. We take into account only links between documents that belong to different domains. During retrieval, after forming the elite set, consisting of the top 1000 retrieved documents, we calculate a new $RSV'$ using a linear combination of the $RSV$ initially computed and the Absorbing Model's authority score $auth_{AM}$:

$$RSV' = a_{content} * RSV + a_{link} * auth_{AM} \quad (1)$$

| Offi cial run | Prec. at 10 | Prec. at 20 | Prec. at 30 | Average Prec. | Features |
|---|---|---|---|---|---|
| uog01ctaialh | 0.1306 | 0.1255 | 0.1218 | 0.1072 | body-anchor-title , $I(n_e)C2$ SAM, SSA |
| uog02ctadh | 0.1143 | 0.1184 | 0.1116 | 0.0979 | body-anchor-title, $I(n_e)C2$ DAM ($\|B\| = 50, \|A\| = 10$), SSA |
| uog03ctadqh | 0.1939 | 0.1612 | 0.1476 | 0.1582 | body-anchor-title, $I(n_e)C2$ DAM ($\|B\| = 50, \|A\| = 10$), DSA |
| uog04cta2dqh | 0.2082 | 0.1704 | 0.1469 | 0.1743 | body-anchor-title , $I(n_e)B2$ DAM ($\|B\| = 50, \|A\| = 10$), DSA |
| uog05tad | 0.2224 | 0.1765 | 0.1565 | 0.1540 | body-anchor-title, $I(n_e)C2$ DAM ($\|B\| = 50, \|A\| = 10$) |

Table 1: Topic distillation offi cial results

The parameters $a_{content}$ and $a_{link}$ were experimentally set to 1 and 0.1 respectively.

The DAM is applied in runs uog02ctadh, uog03ctadqh, uog04cta2dqh and uog05tad for the topic distillation task. After the elite set of documents is ranked according to the matching function used, the $RSVs$ calculated are used as prior probabilities for the initialisation of the DAM. This dynamic link analysis is applied on the set $B$ of the top ranked documents, ignoring the outlinks from the documents of the set $A$, where $A \subset B$. The values used in the official results for the sizes of sets $B$ and $A$ are respectively 50 and 10.

The spreading activation mechanism is applied for the topic distillation task, either statically for runs uog01ctaialh and uog02ctadh, or in a dynamic mode for runs uog03ctadqh and uog04cta2dqh, as a filter for re-ranking the results. The static version (SSA) consists of forming for each document $d$ in the elite set, the set $S$ of documents that are linked by it and that are placed deeper in the hierarchy of documents within the same site. The final $RSV'$ for document $d$ is then calculated by using the following formula:

$$RSV'_d = RSV_d + \beta * \sum_{s \in S} RSV_s \tag{2}$$

The parameter $\beta$ was experimentally set to 0.1 .

The dynamic version (DSA) of the spreading activation mechanism replaces the parameter $\beta$ with the query scope, that is a measure of the specificity of the query. The Equation 2 is then modified as follows:

$$RSV'_d = RSV_d + query_{scope} * \sum_{s \in S} RSV_s \tag{3}$$

As mentioned above in Section 3.1, for the run uog05tad, we use a different approach. Having augmented the inverted file with information on whether the terms belong to the body of a document or to its anchor text, before the DAM is applied, we remove from the results those documents for which no query terms occur in the associated anchor text. The obtained official results for the topic distillation task are given in Table 1.

### 3.3 Named page finding

For the named page finding task, proximity search is used only for the run uog08ctap. After retrieval is performed and the elite set is formed, proximity search is applied and the original $RSV$ of documents in which the query occurs as a phrase is multiplied by a parameter $\gamma$, as shown in the following equation:

$$RSV' = \gamma * RSV \tag{4}$$

| Official run | Average Reciprocal Precision | Named pages in top 10 | Named pages not found | Features |
|---|---|---|---|---|
| uog06c | 0.552 | 107 (71.3%) | 23 (15.3%) | body only, $I(n_e)C2$ |
| uog07cta | 0.654 | 128 (85.3%) | 14 (9.3%) | body-anchor-title, $I(n_e)C2$ |
| uog08ctap | 0.516 | 114 (76.0%) | 18 (12.0%) | body-anchor-title, $I(n_e)C2$<br>Proximity search |
| uog09cta2 | 0.643 | 127 (84.7%) | 12 (8.0%) | body-anchor-title, $I(n_e)B2$ |
| uog10ctad | 0.651 | 128 (85.3%) | 14 (9.3%) | body-anchor-title, $I(n_e)C2$<br>DAM ($\|B\| = 10, A = \|5\|$) |

Table 2: Named page finding official results

| Unofficial run | Prec. at 10 | Prec. at 20 | Prec. at 30 | Average Prec. | Features |
|---|---|---|---|---|---|
| unof01cta | 0.2082 | 0.1714 | 0.1537 | 0.1685 | body-anchor-title, $I(n_e)C2$ |
| unof02c | 0.2122 | 0.1806 | 0.1619 | 0.1668 | body only, $I(n_e)C2$ |
| unof03c | 0.2694 | 0.1929 | 0.1680 | 0.2041 | body only, $PL2$ |
| unof04cd | 0.2776 | 0.1969 | 0.1687 | 0.2047 | body only, $PL2$<br>DAM ($\|B\| = 50, \|A\| = 20$) |
| unof05cdpr | 0.1939 | 0.1612 | 0.1463 | 0.1335 | body only, $PL2$<br>Dynamic PageRank<br>($\|B\| = 50, \|A\| = 20$) |
| unof06cqe | 0.2388 | 0.1888 | 0.1653 | 0.2021 | body only, $PL2$<br>Query Expansion |

Table 3: Topic distillation unofficial results

For the run uog08ctap the parameter $\gamma$ was set to 1.3.

For the run uog10ctad, the DAM is applied as described in Section 2.3. The sizes of the sets $B$ and $A$ were experimentally chosen to be 10 and 5. They are smaller than the corresponding values chosen for the topic distillation task, reflecting the fact that there is only one, or two at most named pages for each query.

The rest three runs, namely uog06c, uog07cta and uog09cta2 are variations of body only, or body and anchor indexing retrieval, using different retrieval methods. Run uog06c uses body only indexing, while runs uog07cta and uog09cta2 use body and anchor indexing. Moreover, runs uog06c and uog07cta use the retrieval method $I(n_e)C2$, while for the run uog09cta2 the method $I(n_e)B2$ was applied. The obtained official results for the named page finding task are given in Table 2.

# 4 Analysis of results

This year, both tasks require a high early precision. The number of relevant documents for the topic distillation task is smaller than the number of relevant documents found for the topic relevance task in TREC 10. The named page finding task differs from the homepage finding task, because the named pages are not necessarily homepages.

For the sake of completeness, once we received the evaluation data from TREC, we ran several new unofficial experiments (see Tables 3 and 4).

Both official and unofficial results obtained from the conducted experiments, show that the content analysis is still the most important/efficient retrieval component. For example, in the topic distillation task, our content-only baseline, unofficial run unof03c, as shown in Table 3, performs better than all our official runs (0.2694 of prec. @10 obtained by run unof03c w.r.t. 0.2224 obtained by

| Unoffi cial run | Average Reciprocal Precision | Named pages in top 10 | Named pages not found | Features |
|---|---|---|---|---|
| unof07ctad | 0.555 | 107 (71.3%) | 22 (14.67%) | body only, $I(n_e)C2$<br>DAM ($\|B\| = 10, \|A\| = 5$) |
| unof08cqe | 0.414 | 93 (62.0%) | 38 (25.3%) | body only, $I(n_e)C2$<br>Query expansion |
| unof09cta | 0.614 | 124 (82.67%) | 11( 7.33%) | body-anchor-title, $PL2$ |

Table 4: Named page finding unoffi cial results

our best run, uog5tad). Note here the application of the weighting scheme $PL2$, which we found to clearly outperform others schemes mentioned in [3].

To have a clear view of the importance of link analysis in topic distillation, we have tuned our link analysis model, the DAM, running several experiments with different values for the sets $B$ and $A$. We have observed a slight improvement of precision at 10 documents and average precision with respect to the content only baseline using the DAM (0.2776 of prec. @10 obtained by run unof04cd w.r.t. 0.2694 obtained by the pure $PL2$ content retrieval baseline, run unof03c, as shown in Table 3). This result was not achieved with the official runs.

Amongst the official runs, uog05tad was the best. Performance difference between the official run uog05tad and the unofficial run unof04cd was due to the fact that in the official run, the use of the anchor and the title text was detrimental, the size of $A$ was too small for a precision @10 and the content retrieval model was different.

We have also compared the DAM to PageRank under the same experimental setting for the topic distillation task. PageRank is applied on the set $B$ of the top $|B|$ documents and the outlinks of the top $|A|$ documents are ignored, where $0 \leq |A| < |B|$. Results show that DAM significantly outperforms PageRank for different values of $|B|$ and $|A|$ (e.g. run unof05cdpr w.r.t. run unof04cd in Table 3). Moreover, PageRank seems to be detrimental for topic distillation (run unof05cdpr w.r.t. run unof02c in Table 3).

We also achieve a slight improvement over the body-only indexing retrieval baseline for the named page finding task (run unof07ctad in Table 4 w.r.t. run uog06c in Table 2) by using our link analysis model, DAM. Although the improvement is marginal, and the sizes of the sets $B$ and $A$ on which the link analysis was applied are smaller than the corresponding sizes for the topic distillation task, this is an indication that our dynamic link analysis model may be applied for both tasks.

An interesting issue that arises from the conducted experiments, concerns the use of anchor text during retrieval. While for the named page finding task, employing anchor text significantly improves precision (run uog07cta w.r.t. run uog06c in Table 2), for the topic distillation task precision decreases (run unof02c w.r.t. run unof01cta in Table 3).

Also, the query-biased spreading activation mechanism proposed seems to significantly improve results over the statically applied spreading activation (run uog03ctadqh w.r.t. run uog02ctadh in Table 1), although the effectiveness is lower than that of our baseline. A possible reason for this is that the spreading activation mechanism was applied on the set of the 1000 top ranked documents, where not all the links contained are useful and most of the documents are not relevant. Therefore, additional refinements are needed, as well as an investigation on the size of the set of documents for which the spreading activation mechanism will prove to be effective. Furthermore, we have noted that query expansion is detrimental for both tasks (run unof06cqe in Table 3 for topic distillation and run unof08cqe in Table 4 for named page finding).

Moreover, the conducted experiments show that the two tasks of the Web track are intrinsically different. Thus, different strategies are needed for each one, since what works best for one task is

not necessary optimal for the other. First, the best results in the two tasks were obtained by using different weighting schemes, i.e. $PL2$ works better for topic distillation (unofficial run `unof03c` w.r.t. unofficial run `unof02c` in Table 3), while $I(n_e)C2$ performs the best in named page finding (official run `uog07cta` w.r.t. unofficial run `unof09cta`). Second, we have proved that while using anchor text improves precision for the named page finding task, it decreases performance in the topic distillation task.

To conclude, for the topic distillation task, both body-only indexing and link analysis without anchors work well, whilst for the named page finding task body and anchor indexing also display promising results. Furthermore, our results show that the potential application and usefulness of the link analysis still has to be explored, and we believe that performance improvement is feasible through refinement and better integration of the content and link analyses. As far as we know, even though content analysis is still a major component for effective retrieval, it is the first time that the results benefit from the application of pure link analysis for both tasks.

## Acknowledgments

This work is funded by a UK Engineering and Physical Sciences Research Council (EPSRC) grant, number GR/R90543/01.

## References

[1] G. Amati, C. Carpineto, and G. Romano. FUB at TREC 10 web track: a probabilistic framework for topic relevance term weighting. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the 10th Text Retrieval Conference TREC 2001*, pages 182–191, Gaithersburg, MD, 2002. NIST Special Pubblication 500-250.

[2] G. Amati and I. Ounis. The absorbing link model for the web. *Manuscript*, 2002.

[3] 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.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[5] D. Hawking and N. Craswell. Overview of the TREC-2001 Web Track, NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001), 2001.

[6] R. Jin and S. Dumais. Probabilistic combination of content and links. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 402–403. ACM Press, 2001.

[7] V. Plachouras and I. Ounis. Query-Biased Combination of Evidence on the Web. Workshop on Mathematical/Formal Methods in Information Retrieval, ACM SIGIR Conference, 2002.

[8] J. Savoy and J. Picard. Retrieval effectiveness on the web. *Information Processing & Management*, 37(4):543–569, 2001.