

# A Utility-oriented Hyperlink Analysis Model for the Web

Vassilis Plachouras  
University of Glasgow  
Glasgow, G12 8QQ, U.K.  
vassilis@dcs.gla.ac.uk

Iadh Ounis  
University of Glasgow  
Glasgow, G12 8QQ, U.K.  
ounis@dcs.gla.ac.uk

Gianni Amati  
Fondazione Ugo Bordoni  
Rome, Italy  
gba@fub.it

## Abstract

*The analysis of hyperlink structure on the Web has been employed for detecting high quality documents. Quality may correspond to the authority of a document, but could also correspond to its utility, that is how well it enables a user to browse its vicinity. We present a hyperlink analysis model, based on modelling the Web graph as an absorbing Markov chain, that can be employed for both authority and utility analysis of documents. The results of experiments for both types of hyperlink analysis underpin the importance of making this distinction. In addition, we provide evidence that support the investigation of more elaborate hyperlink analysis methods on a query-by-query basis.*

## 1. Introduction

The analysis of hyperlink structure of Web documents has been employed in order to discover documents of high quality on the Web. The term quality, as used here, may have more than one interpretation. The first is that of authority: documents that are pointed by many other documents, or by other quality documents are considered to be more authoritative on their topic. This is the approach taken in PageRank and its modifications [4, 10, 15], where the quality of a document depends on its incoming links.

A more refined approach is employed by Kleinberg in the definition of HITS [11], where documents may be authorities and hubs at the same time. In this context, a good authority is pointed by many good hubs, while a good hub should point to many good authorities. The quality of a document as a hub denotes how useful it is for a user who wants to discover the most authoritative documents about a topic. However, apart from HITS and its extensions [13, 6, 3, 5], this distinction has been overlooked and methods for hyper-

link analysis focus mostly on predicting the authority of documents.

We propose a model for hyperlink analysis, namely the *Absorbing Model*, which may be used to measure either the authority of a document, or its utility, that is how well it enables a user to browse its vicinity. Based on modelling the Web graph as a Markov chain, we take a different approach from PageRank, where all documents are linked to each other, in order to get an ergodic Markov chain. In our case, we introduce a set of new states in the Markov chain, uniquely associated with each state in the original Markov chain. The implication of this transformation is that the resulting Markov chain does not possess a stationary probability distribution, and as a result, the prior probabilities of documents affect the hyperlink analysis scores.

Because the concepts of authority and utility are different than that of relevance, we need to combine evidence from both content and hyperlink analysis, in order to achieve effective retrieval [3]. For the combination of evidence, there are different approaches, ranging from a simple weighted sum to more elaborate models, such as the belief network model [17]. We choose a simple and effective formula, the Cobb-Douglas utility function, which corresponds to a weighted product of the different evidence.

For the evaluation of the model, we experiment using the .GOV, a standard TREC Web test collection. We employ the topics and relevance assessments from the topic distillation task of TREC 2002 [7]. This particular task was designed to test the usefulness of hyperlink analysis techniques in finding good entry points to topics. We evaluate both authority and utility interpretations of the Absorbing Model. Especially for the latter, we provide results from an extensive experiment, where the ideal performance of the model is obtained from a set of runs with varying parameters.

The remainder of this paper is organised in the following way. Section 2 presents the basic properties of Markov chains and introduces the Absorbing Model. In

Section 3, we focus on the authority interpretation of the model, while Section 4 introduces the utility interpretation of the model. We report the results of an extensive experiment with the utility Absorbing Model in Section 5. Section 6 provides the conclusions drawn from this work and some interesting points for future work.

## 2. The Absorbing Model

The Web graph can be modelled as a Markov chain, where the probability of accessing a document, while performing a random walk, can be used to indicate the document's authority, or utility. In this section, we will give some of the basic definitions for Markov chains, where the notation and the terminology introduced are similar to that used by Feller [8].

### 2.1. Markov chains

Each document  $d_i$  is considered as an alternative state with prior probability  $p_i$  of being retrieved by the system. In addition, for every pair of documents  $(d_i, d_j)$ , we assign a transition probability  $p_{ij}$ . Both prior and transition probabilities should satisfy the condition of a probability space:

$$\sum_i p_i = 1 \quad (1)$$

$$\sum_j p_{ij} = 1 \quad (2)$$

Moreover, if the probability  $p_{ij}^n$  of reaching document  $d_j$  from document  $d_i$  by a  $n$ -step random walk is greater than zero for some  $n$ , we say that  $d_j$  is *reachable* from  $d_i$ . A set of states is said to be *closed* if any state inside the set can reach all and only all other states in the set. The states in a closed set are called *persistent* or *recurrent* states, since a random walk, starting from the state  $d_i$  and terminating at state  $d_j$ , can be infinitely extended to pass through  $d_i$  again. Indeed, from the definition of the closed set, the probability  $p_{ji}^m$  is greater than zero for some  $m$ . If a single state forms a closed set, then it is called *absorbing*, since a random walk that reaches this state cannot visit any other states anymore. A state which is not in any closed set is called *transient*. Such a state must reach at least one state in a closed set, and consequently, there is a random walk, starting from the transient state  $d_i$ , that cannot be infinitely extended to pass through  $d_i$  again.

It can be shown that all Markov chains can be decomposed in a unique manner into non-overlapping closed sets  $C_1, C_2, \dots$  and a set  $T$  that contains all and

only all the transient states of the Markov chain. If this decomposition results into a single closed set  $C$ , then the Markov chain is called *irreducible*. In this case, the Markov chain possesses a *stationary*, or *invariant* distribution, which is independent of the prior probabilities.

### 2.2. Definition of the Absorbing Model

Modelling the Web graph as a Markov chain has two main implications:

1. There are Web documents that do not have any hyperlinks to other documents. In this case, the probability space condition (2) for the transition probabilities is not satisfied and the definition of a Markov chain becomes problematic [8].
2. Even if every document has at least one outgoing link, there are states in the Markov chain, namely the *transient* states, for which the limit of the probability of accessing the corresponding state is zero. In other words, although a random walk may start from a specific document, it is not possible to ever extend the random walk in order to pass from the initial document again.

There are two alternatives in order to overcome these two implications of modelling directly the Web graph as a Markov chain:

1. All states are linked by assigning a new transition probability  $p_{ij}^* \neq 0$  in a suitable way. In this case there are no more transient states and all states receive a positive authority score. This approach is used in PageRank, where the assumed random surfer may randomly jump with a finite probability to any Web document.
2. The original graph  $G$  is extended to a new graph  $G^*$ , where the new states of  $G^*$  are all and only all the persistent states. The scores relative to all states of the original graph, whether transient or not, will be uniquely associated to the scores of the new states of  $G^*$ .

The Absorbing Model is defined by following the second alternative. We project the original graph  $G$  onto a new graph  $G^*$  whose decomposition is made up of a set of transient states  $T = G$  and a set  $C_1, \dots, C_n$  of absorbing states, that is a set of singular closed sets. The state  $C_i$  is called the *clone* of state  $d_i$  of the original graph  $G$ . Any state in  $G$  has access only to its corresponding clone, but not to other clones. Since the clones are absorbing states, they do not have access to any state except to themselves. The Absorbing Model is introduced formally as follows:

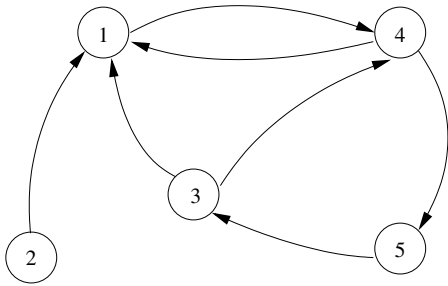
**Definition 1** Let  $G = (D, R)$  be the graph consisting of the set  $D$  of the  $N$  documents  $d_i$  and the binary accessibility relation  $R(d_i, d_j) = 1$  if there is a hyperlink from  $d_i$  to  $d_j$  and 0 otherwise. The graph  $G$  is extended by introducing  $N$  additional states  $d_{N+i}, i = 1, \dots, N$ , called the clone nodes. These additional nodes are denoted as:  $d_{N+i} = d_i^*$  and the accessibility relation  $R$  is extended in the following way:

$$\begin{aligned} R(d_i^*, d) &= R(d, d_i^*) = 0, d \neq d_i^*, i = 1, \dots, N \text{ except for:} \\ R(d_i, d_i^*) &= 1 \\ R(d_i^*, d_i^*) &= 1 \end{aligned}$$

The transition probability  $p_{ij}$  from state  $d_i$  to state  $d_j$  is:

$$p_{ij} = \frac{R(d_i, d_j)}{|\{d_j : R(d_i, d_j) = 1\}|} \quad (3)$$

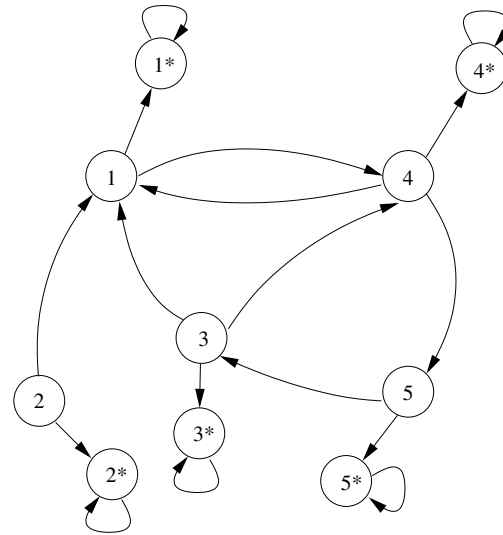
where the denominator stands for the number of the possible transitions from state  $d_i$ .



**Figure 1. The Markov Chain representing the web graph**

Before continuing, we will give an example that illustrates the transformation of the graph. In Figure 1, a graph that represents a part of the Web is shown. According to the definitions given for Markov chains, states 1, 3, 4 and 5 form a closed set and they are persistent states. State 2 is a transient state. Therefore the corresponding Markov chain is not irreducible, as it can be decomposed to a non-empty set of transient states and one set of persistent states. Figure 2 shows the same graph, transformed according to the definition of Absorbing Model. In this case, the states 1 to 5 become transient and the only persistent states are the newly introduced states  $1^*$  to  $5^*$ . In addition, we should note that the introduced transformation results in removing any absorbing states from the original Web graph, as there are no closed sets composed by any of the original states.

Hence, with the introduction of the clone nodes, all the original states  $d_j, j = 1, \dots, N$  become transient,



**Figure 2. The extended Markov Chain including the clone states**

while all the clone states  $d_j^*, j = 1, \dots, N$  are the only persistent states. In other words, for the states in the original Markov chain we have:

$$p_{jk}^n \rightarrow 0, k = 1, \dots, N \quad (4)$$

while for the clone states we have:

$$p_{jk}^n \rightarrow u_{jk}, k = N + 1, \dots, 2N \quad (5)$$

where  $u_{jk}$  stands for the probability that a random walk starting from state  $d_j$  will pass through state  $d_k$ . We define the Absorbing Model score of a state  $d_k$  to be given by the unconditional probability of reaching its clone state  $d_k^*$ :

$$\sum_j p_j u_{jk^*} \quad (6)$$

where  $k^* = k + N$  and  $k = 1, \dots, N$ .

Intuitively, the Absorbing Model score measures the probability of a user being “absorbed” by a Web document, while he is browsing other documents in its vicinity. This probability depends on both incoming and outgoing links:

1. If a document has many outgoing links, then its Absorbing Model score is low, while if it has few outgoing links, it is more probable that its Absorbing Model score will be higher. Therefore, the low values of the Absorbing Model score can be considered as evidence of utility (or hub quality) for documents.

2. Documents with a significant number of incoming links, have a high Absorbing Model score, while documents without incoming links have a lower score. Therefore, the higher values of the Absorbing Model score can be considered as evidence of authority for documents.

Before continuing, we would like to point two main qualitative differences between the Absorbing Model and PageRank. First, while in PageRank the scores depend mainly on the quality of the incoming links of a document, in the Absorbing Model the document's score is affected by its outgoing links. Thus, it allows us to introduce and quantify the concept of utility, as it will be described in Section 4.

The second difference is that PageRank scores correspond to the stationary probability distribution of the Markov chain resulting from the Web graph after adding a link between every pair of documents. On the other hand, the Absorbing Model does not possess a stationary distribution, and therefore, the Absorbing Model scores depend on the prior probabilities of the documents. Depending on the way the prior probabilities are defined, we may introduce different extensions of the model. For example, the use of the content retrieval scores as the prior probabilities results into a simple and principled way to combine dynamically content and link analysis [1], similarly to the extensions of HITS [3]. On the other hand, if the prior probabilities are defined independently of the content retrieval, as we will see in the next sections, we can compute the Absorbing Model scores offline, as in the case of PageRank.

In this paper, we focus on the latter approach for defining the prior probabilities, and introduce the authority-oriented (Section 3) and the utility-oriented (Section 4) interpretations of the model.

### 3. The Static Absorbing Model

From the possible ways to define the prior probabilities, such as the url type, or the document's length [12], one is to assume that they are uniformly distributed. This approach reflects the concept that all the documents are equally likely to be retrieved, without taking into account any of their specific characteristics. Consequently, the prior probabilities are defined as follows:

**Definition 2** *Static mode priors: the prior probability that the document  $d_k$  is retrieved is uniformly distributed over all the documents:*

$$p_k = \frac{1}{2N} \quad (7)$$

where the number  $2N$  refers to the total number of states in the new graph, that is the total number of documents, plus an equal number of the corresponding clone states.

When we employ the static mode priors, the Absorbing Model score  $s(d_j)$  of a document  $d_j$  is given from (6) and (7) as follows:

$$s(d_j) = \sum_i p_i u_{ij^*} = \sum_i \frac{1}{2N} u_{ij^*} \propto \sum_i u_{ij^*} \quad (8)$$

In other words, the Absorbing Model score  $s(d_j)$  for a document  $d_j$  is the probability of accessing its clone node  $d_j^*$  by performing a random walk, starting from any state with equal probability. The interpretation of this score is derived in a straightforward manner from the intuitive description of the Absorbing Model in Section 2: a document has a high Absorbing Model score if there are many paths leading to it. As a result, a random user would be absorbed by the document, while he would be browsing the documents in its vicinity. Highly authoritative documents are favoured by this approach, and they are expected to have a higher Absorbing Model score.

In order to combine the Absorbing Model score with the content analysis score, we employ a Cobb-Douglas utility function, as follows:

$$U = C^a \cdot L^b, \quad a + b = 2 \quad (9)$$

This utility function has been applied successfully to combine different sources of utility, such as labour and capital in the context of economics. The exponents  $a$  and  $b$  are parameters that regulate the importance of each of the components in the combination, and by definition they sum up to 2.

In our case, we combine the content analysis score  $s(d_j|q)$  for query  $q$  and the Absorbing Model score  $s(d_j)$ , using equal values for the exponents  $a = b = 1$ , and the final score for a document  $d_i$  is given as follows:

$$U_i = s(d_j|q) \cdot s(d_j) \quad (10)$$

We refer to this method as the Static Absorbing Model (SAM).

To test this approach, we have performed experiments using a standard Web test collection, namely the .GOV, a recent crawl from the .gov domain used for the Web track of TREC 2002 [7]. We used the 49 queries and the corresponding relevance assessments from the topic distillation task of the Web track of the same TREC, a task focused on finding useful resources about a topic.

For the content analysis, we employed the *BM25* weighing function [16] and two weighting functions

	Aver. Prec.	Prec. at 5	Prec. at 10
<i>BM25</i>	0.1904	0.2939	0.2429
<i>SAM<sub>BM25</sub></i>	0.0020	0.0082	0.0041
<i>PR<sub>BM25</sub></i>	0.0032	0.0041	0.0204
<i>PL2</i>	0.2031	0.3020	0.2694
<i>SAM<sub>PL2</sub></i>	0.0031	0.0082	0.0041
<i>PR<sub>PL2</sub></i>	0.0039	0.0163	0.0245
<i>I(n<sub>e</sub>)B2</i>	0.1994	0.3020	0.2408
<i>SAM<sub>I(n<sub>e</sub>)B2</sub></i>	0.0027	0.0082	0.0041
<i>PR<sub>I(n<sub>e</sub>)B2</sub></i>	0.0039	0.0122	0.0245

**Table 1. Authority-oriented experiments with Static Absorbing Model and PageRank**

from the probabilistic framework proposed by Amati and Van Rijsbergen [2], namely the functions  $I(n_e)B2$  and  $PL2$ . For the hyperlink analysis, the Absorbing Model scores were computed during indexing, employing all the hyperlinks in the collection, and they were normalised by dividing by the maximum of the scores. We should note that the computational overhead due to the introduction of the clone states was insignificant. In addition, we ran experiments using PageRank (denoted by PR in Table 1) instead of the Absorbing Model, while all the other settings were the same.

Results from Table 1 show that for both SAM and PageRank, the authority oriented approach is not effective for retrieval on the specific collection, independently of the weighing function used. However, a close look at the collection and the task could suggest that this authority-oriented approach may not be suitable for application on a collection where all sources are of high quality and authoritative. In the collection under consideration, the quality derives from the authority of the authors and the hyperlinks that point to the documents in the collection from external documents. The latter set of hyperlinks is not part of the collection and therefore cannot be used to leverage authority.

In addition, it could be the case that authority-oriented analysis may not be suitable for applying on a per document basis, but may behave differently when applied on aggregates of documents. An analogous method is employed in the field of citation analysis, where the impact factor for journals is used to denote the importance of specific journals [9]. The impact factor is not computed for single papers, but for aggregates of papers, which are published in the same journal. However, we should note that it is not straight-

<sup>1</sup> The indices of SAM and PR in Tables 1 and 2 denote the weighting method used for the content analysis.

forward to relate the fields of citation and hyperlink analysis, since the motivations for adding citations in a scientific paper are different than the motivations for adding a hyperlink to a Web document [18].

#### 4. The Static Utility Absorbing Model

In this section, we focus on a different approach to hyperlink analysis, where we do not consider authority, but utility. In our context, the term utility corresponds to the concept of how well a document enables a user to browse its vicinity. For example, a document with few outgoing links, or with outgoing links to irrelevant documents, is not particularly helpful in this sense. On the other hand, a document with a high number of outgoing links to relevant documents may be seen as a useful resource. In order to make a fair comparison to SAM, for the remainder of the paper, the concept of utility will be related to the number of outgoing links from a document, as described below.

We modify the Static Absorbing Model as follows. The prior probabilities are assigned to documents in exactly the same way as in the case of the Static Absorbing Model, but instead of using the Absorbing Model score  $s(d_j)$  for document  $d_j$ , we employ its informative content  $-\log_2(s(d_j))$  [14]. As already mentioned in Section 2, the Absorbing Model score of a document depends on both incoming and outgoing links of the document. In addition, the probability of accessing the clone node of a document is lower for documents with a higher number of outgoing links. For this reason, we adopt the informative content of the Absorbing Model score, which measures the importance of encountering a document with a low Absorbing Model score.

Again, for the combination of evidence, we employ the Cobb-Douglas utility function, as it is introduced in (9) with exponents  $a = b = 1$ , replacing the Absorbing Model score by its informative content:

$$U_i = s(d_j|q) \cdot (-\log(s(d_j))) \quad (11)$$

We refer to this method as the Static Utility Absorbing Model (SUAM).

We should note that the use of the informative content of the PageRank scores is not intuitive, since PageRank is meant to measure authority. Therefore, the low PageRank scores suggest nothing about the utility of the corresponding documents, but only about their low authority. Hence, it is not appropriate to make a direct comparison between SUAM and PageRank.

For testing the effectiveness of SUAM, we experiment in the setting described in Section 3. As it can be

	Aver. Prec.	Prec. at 5	Prec. at 10
<i>BM25</i>	0.1904	0.2939	0.2429
SUAM <sub><i>BM25</i></sub>	0.1850	0.2980	0.2286
<i>PL2</i>	0.2031	0.3020	0.2694
SUAM <sub><i>PL2</i></sub>	0.2046	0.3061	0.2490
<i>I(n<sub>e</sub>)B2</i>	0.1994	0.3020	0.2408
SUAM <sub><i>I(n<sub>e</sub>)B2</i></sub>	0.1957	0.3306	0.2469

**Table 2. Static Utility Absorbing Model results**

seen from Table 2, the results are at the level of the content only baseline. While the average precision decreases marginally, the precision at 5 documents shows an improvement for all the weighting functions used for content analysis. Precision at 10 documents is improved only when the weighting function *I(n<sub>e</sub>)B2* is employed. The decrease of average precision is not particularly important since the official measure for the topic distillation task of TREC is the precision at 10 documents. More generally, in the context of Web Information Retrieval, we are interested more in precision amongst the top retrieved documents, since users tend to browse only a few top ranked documents. For the rest of the paper, we will only use the weighting function *I(n<sub>e</sub>)B2*, as it is the one resulting into the highest improvement after the combination with hyperlink analysis.

Table 3 presents a query-by-query analysis of the results. The first column refers to the measure used for comparisons between the *I(n<sub>e</sub>)B2* and the SUAM<sub>*I(n<sub>e</sub>)B2*</sub> experiments: average precision and precision at 5 and 10 documents. The next three columns correspond to the number of queries for which we observed an improvement (+), a loss of precision (-), or where the results were the same (=). The last column presents the resulting *p* values from the Wilcoxon's signed ranks test for paired samples. All three measures were tested for statistical significance, but the result did not show that SUAM<sub>*I(n<sub>e</sub>)B2*</sub> resulted into a significant improvement. There was an improvement for precision at 10 documents for 11 queries, while a decrease was observed for 10 queries. As for precision at 5 documents, there was an improvement for 13 queries and a decrease for 6. If we consider average precision, an increase is obtained for 24 queries, for a single query the average precision is the same and for the rest of the queries it drops.

Overall, there is a tendency towards increasing the precision of the top ranked documents. The observed improvement, although not significant, indicates that the utility-oriented link analysis is more appropriate for

Measure	+	-	=	<i>p</i> (Signed ranks test)
Aver. Prec	24	24	1	0.5417
Prec. at 5	13	6	30	0.3284
Prec. at 10	11	10	28	0.8202

**Table 3. Query-by-query analysis: *I(n<sub>e</sub>)B2* versus SUAM<sub>*I(n<sub>e</sub>)B2*</sub>**

the topic distillation task under consideration, since a useful resource on a topic is expected to point to other relevant documents on the same topic. Comparing the results of the utility-oriented SUAM to the authority-oriented SAM, described in Section 3, we can say that the former is more effective and reliable than the latter, in this TREC task.

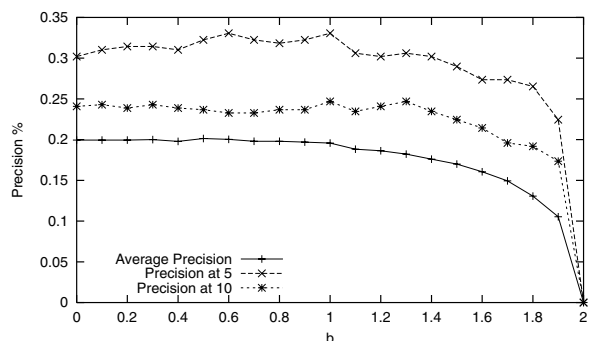
## 5. Extended experiment with the Static Utility Absorbing Model

In order to further examine the Static Utility Absorbing Model, we investigate the effect of adjusting the parameters *a* and *b* in the utility function (9). The exponents represent the relative importance of each of the components used in the combination of evidence. We should note that effectively we introduce only one parameter in the model, because the sum of the exponents should be constant, i.e. equal to 2.

We have conducted an experiment in which we set the exponents *a* and *b* to values between 0 and 2 in steps of 0.1. In Figure 3, the average precision, and the precision at 5 and 10 documents are presented as a function of the exponent *b*. It should be noted that the precision of the *I(n<sub>e</sub>)B2* content-only baseline corresponds to the points for *b* = 0 and that the exponent for the content-based module is *a* = 2 - *b*. We can see that our model is relatively stable across a wide range of values for *b* and its performance is decreasing rapidly for values of *b* larger than 1.3.

More importantly, in the range of *b* values where the model's performance is stable, there are specific points where improvement over the content-only baseline *I(n<sub>e</sub>)B2* is observed. For example for *b* = 0.6, we have 0.3306 average precision at 5 documents, which according to Wilcoxon's signed ranks test for paired samples, is significant with *p* = 0.0492. The same value for precision at 5 documents is observed for *b* = 1.0 but this does not prove to be significant according to the same statistical test. In addition, improvements over the content-only baseline's precision at 10 documents are observed for specific values of *b*. At the points where *b* = 1.0, or *b* = 1.3 we have 0.2469 precision at 10 docu-

ments with respect to 0.2408 of the content only baseline, although this is not significantly different.



**Figure 3. Precision in relation with the values of the exponents**

So far, we have considered applying the same values for  $a$  and  $b$  for ranking the results of all the queries under consideration. However, if we look into the best values for the parameters in a query-by-query basis, we observe that two groups of queries may be identified. The first group of queries consists of those that do not benefit from the application of SUAM. For these queries, the retrieval effectiveness when applying SUAM is either stable, or drops (for 6 out of the 32 queries in this group, no relevant documents were retrieved by our system). The other group of queries consists of those queries where the application of SUAM increases precision. If we use precision at 10 documents for grouping the queries, we find that the first group consists of 32 queries and the second one consists of 17 queries.

Since we have conducted the experiments with all the possible combinations of the exponents, we can see what would be the effectiveness of this model under the assumption that we have a mechanism for predicting the best values for the parameters  $a$  and  $b$  on a query-by-query basis for the corresponding measure of retrieval effectiveness. Table 4 summarises our findings. For example, in the ideal case, where we could find the most appropriate values for  $a$  and  $b$  in order to maximise average precision (Maximum1), the resulting average precision would be 0.2201 (significant with  $p < 10^{-5}$ ), precision at 5 documents would be 0.3674 (significant with  $p = 0.0039$ ) and precision at 10 documents would be 0.2857 (significant with  $p = 0.0081$ ).

Alternatively, we tested a more realistic assumption that there is a mechanism for approximating the best values for parameters  $a$  and  $b$ . Even if such a mechanism returned the values for the parameters that would

Measure	Aver. Prec.	Prec@5	Prec@10
$I(n_e)B2$	0.1994	0.3020	0.2408
Maximum1	0.2201	0.3674	0.2857
Maximum3	0.2086	0.3612	0.2531

**Table 4. Comparison between  $I(n_e)B2$  and the ideal cases**

correspond to just the third best average precision per query, precision amongst the top ranked documents would still improve considerably (see Maximum3 in Table 4). More specifically, using Wilcoxon's signed ranks test, we can see that both average precision and precision at 5 documents improve significantly with  $p$  equal to 0.0078 and 0.0093 respectively. However, according to the same statistical test, although precision at 10 documents is improved over the baseline, this difference is not significant ( $p = 0.4212$ ).

We should note that maximising average precision does not guarantee that precision at 5, or 10 documents will be maximised, but it is highly likely that they will be higher than the corresponding results returned by  $I(n_e)B2$ . For example, if we maximised the average precision, then precision at 10 documents would be 0.2857, while if we aimed at maximising precision at 10, then the precision obtained would be 0.2939. In the same way, maximising average precision results into precision at 5 documents equal to 0.3674, while if we chose to maximise precision at 5 documents, we would get a maximum of 0.4041. The values of the parameters that maximise the average precision result into maximum precision at 10 documents for 40 queries, and into maximum precision at 5 documents for 43 out of the 49 queries respectively. These results show the high correlation between the average precision and the precision at 5 and 10 documents.

Based on the results obtained, we can evaluate realistic methodologies for predicting the values for the same parameters. Indeed, a decision mechanism resulting into precision at 10 documents close to the one obtained under the assumption of employing an ideal mechanism, should assign to parameters  $a$  and  $b$  values close to the optimal ones.

## 6. Conclusions

In this paper, we have presented the Absorbing Model for utilising evidence from the hyperlinks between Web documents to improve Web Information Retrieval effectiveness. Two different interpretations of the Absorbing Model are suggested in order to handle the different types of hyperlink analysis. For an

authority-oriented hyperlink analysis we employ the Static Absorbing Model, that provides an indication of the authority of documents. For a utility-oriented analysis of the hyperlink structure, we propose the Static Utility Absorbing Model, which gives scores to documents according to how well they enable users to browse their vicinity. In both models the computation of the Absorbing Model scores is performed offline and the combination of evidence from the content and hyperlink analysis is achieved by employing the Cobb-Douglas utility function. In the future, we will evaluate a dynamic version of the Absorbing Model, where the prior probabilities will correspond to the content retrieval scores [1].

We have performed experiments with both approaches, using the .GOV Web test collection. For the Static Absorbing Model, more investigation is needed in order to find the reasons that make it inappropriate for this test collection. On the other hand, the Static Utility Absorbing Model is stable and improves precision among the top ranked documents for the same collection. This contrast underpins the difference between the two hyperlink structure analysis approaches. As the experiments suggest, the utility of a document, in terms of how well it enables a user to browse its vicinity, is more effective than its authority in the context of the test collection used. Additional data, such as the topic distillation queries of TREC 2003, will help us to check the validity of these conclusions in different settings.

In addition, we have shown that in the ideal case where the best values for the parameter  $b$  of the model could be chosen automatically, the retrieval effectiveness would be significantly better for the specific test collection. This result is important, because it shows the potential effectiveness of hyperlink structure analysis in the context of TREC-like experiments. In addition, it provides us with maximum values of precision for the specific collection and query set when our approach is applied, and therefore, it can be used to evaluate models for predicting appropriate values for  $b$ . We are currently working on such models, while a point of our future research will be an investigation of methods for the application of the authority-oriented hyperlink analysis.

## ACKNOWLEDGEMENTS

This work 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>). We would like to thank Tassos Tombros and Professor Keith van Rijsbergen for discussing this work and providing useful comments. We would also like to thank Professor Peter Ingwersen for his useful suggestions.

## References

- [1] G. Amati, I. Ounis, and V. Plachouras. The dynamic absorbing model for the web. Technical Report TR-2003-137, Department of Computing Science, University of Glasgow, 2003.
- [2] 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.
- [3] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111. ACM Press, 1998.
- [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] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In L. M. Haas and A. Tiwary, editors, *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*, pages 307–318, Seattle, US, 1998. ACM Press, New York, US.
- [6] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proceedings of the 17th International Conference on Machine Learning*. ACM Press, 2000.
- [7] N. Craswell and D. Hawking. Overview of the TREC-2002 Web Track. In *NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002)*, pages 86–93, 2002.
- [8] W. Feller. *An Introduction to Probability Theory and its Applications, volume 1, 2nd edition*. John Wiley and Sons, 1957.
- [9] E. Garfield. Citation analysis as a tool in journal evaluation. *Science* 178:471–479, 1972.
- [10] T. H. Haveliwala. Topic-Sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.
- [11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [12] 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.
- [13] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):387–401, 2000.

- [14] K. Popper. *The Logic of Scientific Discovery*. Hutchinson & Co., London, 1959.
- [15] M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems 14*, 2002.
- [16] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer-Verlag New York, Inc., 1994.
- [17] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, and N. Ziviani. Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103. ACM Press, 2000.
- [18] M. Thelwall. What is this link doing here? beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8, 2003.