Journal of Web Engineering, Vol. 0, No. 0 (2003) 000–000

The Static Absorbing Model for the Web[a]

Vassilis Plachouras
University of Glasgow
Glasgow G12 8QQ UK
vassilis@dcs.gla.ac.uk

Iadh Ounis
University of Glasgow
Glasgow G12 8QQ UK
ounis@dcs.gla.ac.uk

Gianni Amati
Fondazione Ugo Bordoni
Rome 00142, Italy
gba@fub.it

Received (received date)
Revised (revised date)

The analysis of hyperlink structure on the Web has been employed for detecting high quality documents. In approaches such as PageRank, the Web graph is modelled as a Markov chain and the quality of a document corresponds to the probability of visiting it during a random walk. However, it is not always straightforward to consider the Web graph as a Markov chain. For example, PageRank introduces a universal document, in order to transform the Web graph to a Markov chain.

In this paper, we present the Absorbing Model, a hyperlink analysis model based on absorbing Markov chains, where the Web graph is transformed by adding one absorbing state for each document. We provide an authority-oriented and a utility-oriented interpretation of the Absorbing Model, and show that the latter is more effective than the authority-oriented model. Thus, we believe that it is quite important to make this distinction between the two types of hyperlink analysis. In addition, we provide evidence that support the investigation of more elaborate hyperlink analysis methods on a query-by-query basis.

Keywords: Web information retrieval, hyperlink analysis, markov chains, absorbing model, combination of evidence
Communicated by: to be filled by the Editorial

1 Introduction

The analysis of hyperlink structure of Web documents has been employed in order to discover documents of high *quality* on the Web. Approaches such as PageRank [1, 2] and related works [3, 4, 5, 6], model the Web graph as a Markov chain and compute the probability of visiting a document during a random walk. The quality of a document depends on both the number of incoming links, and the quality of the documents that point to it. For example,

[a]This paper is an extended version of the paper 'A Utility-Oriented Hyperlink Analysis Model for the Web', which appeared in the proceedings of the First Latin American Web Congress, 2003.

if a trusted site, such as www.yahoo.com, links to a document, then this link carries more importance than a link from a random document. PageRank is independent of the queries and therefore, we can compute its output once on the whole Web graph. Therefore, PageRank scores can be employed in query time, without significant overhead.

However, it is not certain that the Web graph can be modelled as a Markov chain in a straightforward manner. For example, we cannot define a Markov chain with states that do not allow the transition to other states. This situation is common on the Web, where documents do not necessarily have any outgoing links. In PageRank, this problem is overcome by introducing a universal document that permits a random transition to any document with a finite probability.

We take a different approach and propose a model for hyperlink analysis, namely the *Absorbing Model*, which can be used in either a query-independent, or a query-dependent way. Based on modelling the Web graph as a Markov chain, we introduce a set of new *absorbing* states, uniquely associated with each state of the original Markov chain. The Absorbing Model score for a document is the probability of visiting the corresponding absorbing state. The implication of this transformation is that the resulting Markov chain does not possess a stationary probability distribution. As a result, the prior probabilities of documents affect the hyperlink analysis scores. This allows for a natural combination of evidence from content and hyperlink analysis, in either a query-independent, or a query-dependent way. Depending on whether the prior probabilities of documents are related to the queries or not, we can define the Dynamic Absorbing Model [7] and the Static Absorbing Model. In this paper, we will focus on the Static Absorbing Model.

We provide two interpretations of the Absorbing Model. First, it can be used to measure the authority of a document, similarly to PageRank. Alternatively, we can employ the Absorbing Model in order to measure the utility of a document, that is how well it enables a user to browse its vicinity. This is similar to Kleinberg's HITS algorithm [8] and related works [9, 10, 11, 12, 13], where documents have two qualities: they can be authorities and hubs. In this paper, we focus on applying the Absorbing Model in a query-independent way, using global hyperlink information [14], where the utility of a document is not related to its relevance, but to the number of its outgoing hyperlinks.

Since the concepts of authority and utility are different from relevance, employing only hyperlink analysis is not sufficient for effective retrieval. Therefore, we need to combine evidence from both content and hyperlink analysis [9]. For the combination of evidence, there are different approaches, ranging from a simple weighted sum to more elaborate models, such as Bayesian network models [15], or belief network models [16]. We choose a simple and effective formula, the Cobb-Douglas utility function, which corresponds to a weighted product of the scores from content and hyperlink analysis.

We evaluate the Absorbing Model in a TREC-like experimental setting. TREC is a yearly forum for the evaluation of large-scale retrieval systems. We use the .GOV Web test collection and the topics and relevance assessments from the topic distillation tasks of TREC11 [17] and TREC12 [18]. We compare the authority-oriented Absorbing Model with PageRank, and evaluate the utility-oriented Absorbing Model. For the latter, we provide results from an experiment, where the ideal performance of the model is obtained from a set of runs with varying parameters. Our results underpin the importance of making the distinction between

the authority and utility-oriented types of hyperlink analysis. We also show that query-biased retrieval approaches can lead to improvements in retrieval effectiveness.

The remainder of this paper is organised in the following way. In Section 2, we present the basic properties of Markov chains and introduce the Absorbing Model. In Section 3, we define and evaluate the authority-oriented Static Absorbing Model. In Section 4, we present and evaluate the utility-oriented Absorbing Model. We report the results from 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 quality. In Sections 2.1, 2.2 and 2.3, we will give some of the basic definitions for Markov chains, and we will define the Absorbing Model in Section 2.4. The notation and the terminology introduced are similar to that used by Feller [19].

2.1 Markov chains

Each document is a possible outcome of the retrieval process. Therefore, we assume that documents are orthogonal, or alternative states d_k, which have a prior probability p_k defined by the system. We associate with each pair of documents (d_i, d_j), a transition probability $p_{ij} = p(d_j|d_i)$ of reaching the document d_j from the document d_i. This conditional probability may be interpreted as the probability of having the document d_j as outcome with the document d_i as evidence.

Both priors and transition probabilities must satisfy the condition of a probability space, which is:

$$\sum_k p_k = 1 \tag{1}$$

$$\sum_j p_{ij} = 1 \tag{2}$$

Condition (2) imposes that each state d_i must have access to at least one state d_j for some j, where it is possible that $i = j$.

It is useful to express the priors as a row vector P and the transition probabilities as a row-by-column matrix M, so that we can have a more compact representation of probabilities for arbitrary sequences of states:

$$P = \left[\; p_k \; \right] \tag{3}$$

$$M = \left[\; p_{ij} \; \right] \tag{4}$$

Then, let M^n be the matrix product rows-into-columns of M with itself n-times

$$M^n = \left[\; p_{ij}^n \; \right] \tag{5}$$

In order to have a Markov chain, the probability of any walk from a state d_i to a state d_j depends only on the probability of the last visited state. In other words, the probability of any sequence of states $(d_1, \ldots, d_n)$ is given by the relation:

$$p(d_1, \ldots, d_n) = p_1 \prod_{i=1}^{n-1} p(d_{i+1}|d_i) \tag{6}$$

where p_1 is the prior probability of document d_1.

In terms of matrices, the element p_{ij}^n of the product M^n corresponds to the probability $p(d_i, \ldots, d_j)$ of reaching the state d_j from d_i by any random walk, or sequence of states $(d_i, \ldots, d_j)$ made up of exactly n states.

If $p_{ij}^n > 0$ for some n, then we say that the state d_j is *reachable* from the state d_i. A set of states $C = \{d_i\}$ is said to be *closed* if any state inside C can reach all and only all other states inside C. 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 ever extended to pass through d_i again. Indeed, from the definition of the closed set, the probability $p_{ji}^m > 0$ 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. A state, which is not in any closed set, is called *transient* and it must reach at least one state in a closed set. Thus, there is a random walk, starting from the transient state d_i, that cannot be ever extended to pass through d_i again.

One of the most useful properties of Markov chains is the decomposition characterisation. It can be shown that all Markov chains can be decomposed in a unique manner into non-overlapping closed sets $C_1, C_2, \ldots, C_n$ and a set T that contains all and only all the transient states of the Markov chain [19]. If this decomposition results in a single closed set C, then the Markov chain is called *irreducible*.

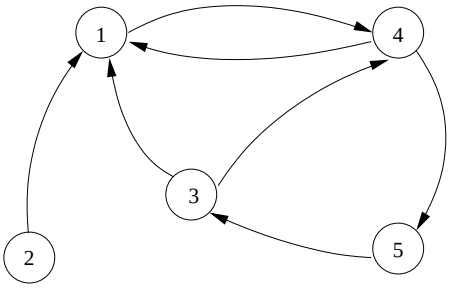

Fig. 1. The Markov Chain representing the Web graph.

We will illustrate the above definitions with the following example. In Figure 1, the directed graph may be seen as the Markov Chain corresponding to few Web documents, with the arcs representing the links between documents and consequently, the transitions between states in the Markov chain. According to the terminology given above for Markov chains, states $1, 3, 4, 5$ form a closed set and they are persistent states. State 2 is a transient state. Therefore this Markov chain is irreducible, as it can be decomposed in a non-empty set of

transient states and a single set of persistent states. Moreover, if the arc from state 5 to state 3 is replaced by an arc from 5 to itself, then state 5 becomes an absorbing state.

2.2 Classification of states

According to Equation (6), the probability of reaching the state d_j from any initial state by any random walk $w = (d_i, \ldots, d_j)$ is given below:

$$\sum_i \sum_w p(d_i, \ldots, d_j) = \sum_i \sum_{n=1}^{\infty} p_i p_{ij}^n = \sum_i p_i \left(\sum_{n=1}^{\infty} p_{ij}^n \right) \tag{7}$$

Therefore, the unconditional probability of reaching a state d_j by any random walk is the limit for $n \to \infty$ of the sum over n of the j-th element of the vector $P \cdot M^n$, which is the product rows-into-columns of the vector P and the matrix M^n.

However, in a Markov chain, the limit $\lim_{n\to\infty} \sum_n p_{ij}^n$ does not always exist, or it can be infinite. The limit does not exist when there is a state d_i such that $p_{ii}^n = 0$ unless n is a multiple of some fixed integer $t > 1$. In this case, the state d_i is called *periodic*. Periodic states are easily handled: if t is the largest integer which makes the state d_i periodic, then it is sufficient to use the probabilities p_{kj}^t as new transition probabilities p'_{kj}. With the new transition probabilities, p'^n_{ii} will be greater than 0 and the periodic states d_j will become aperiodic. Hence, we may assume that all states in a Markov chain are aperiodic [19].

Recurrent states in a finite Markov chain have the limit of p_{ij}^n greater than 0 if the state d_j is reachable from d_i, while for all transient states this limit is 0:

$$\lim_{n\to\infty} p_{ij}^n = 0 \text{ if } d_j \text{ is transient} \tag{8}$$

$$\lim_{n\to\infty} p_{ij}^n > 0 \text{ if } d_j \text{ is persistent and } d_j \text{ is reachable from } d_i \tag{9}$$

In an irreducible finite Markov chain, all nodes are persistent and the probability of reaching them from an arbitrary node of the graph is positive. In other words, $\lim_{n\to\infty} p_{ij}^n > 0$ and $\lim_{n\to\infty} p_{ij}^n = \lim_{n\to\infty} p_{kj}^n = u_j$ for all i and k. Due to this property, an irreducible Markov chain possesses an invariant distribution, that is a distribution u_k such that:

$$u_j = \sum_i u_i p_{ij} \qquad \text{and} \qquad u_j = \lim_{n\to\infty} p_{ij}^n \tag{10}$$

In the case of irreducible Markov chains, the vector P of prior probabilities does not affect the unconditional probability of entering an arbitrary state, since all rows are identical in the limit matrix of M^n. Indeed:

$$\lim_{n\to\infty} \sum_i p_i p_{ij}^n = \lim_{n\to\infty} \sum_i p_i p_{kj}^n = \lim_{n\to\infty} {p_{kj}}^n \sum_i p_i = u_j \left(\sum_i p_i \right) = u_j \tag{11}$$

Because of this property, the probability distribution u_j in a irreducible Markov chain is called *invariant* or *stationary* distribution.

If the distribution $\lim_{n\to\infty} \sum_i p_i p_{ij}^n$ is taken to assign weights to the nodes, then it is equivalent to the invariant distribution u_j in the case that the Markov chain is irreducible. More generally, if the Markov chain is not irreducible or does not possess an invariant distribution, then $\lim_{n\to\infty} \sum_i p_i p_{ij}^n$ can be still used to define the distribution of the node weights. However, it will depend on the prior distribution p_i.

2.3 Modelling the hyperlinks of the Web

In this section we formally present how Markov chains can be applied to model hyperlinks on the Web.

Let R be the binary accessibility relation between the set of documents, namely $R(d_i, d_j) = 1$, if there is a hyperlink from document d_i to document d_j, and 0 otherwise.

Let $o(i)$ be the number of documents d_j which are accessible from d_i:

$$o(i) = |\{d_j : R(i,j) = 1\}| \tag{12}$$

The probability p_{ij} of a transition from document d_i to document d_j is defined as follows:

$$p_{ij} = \frac{R(i,j)}{o(i)} \tag{13}$$

If we model the Web graph with a stochastic matrix defined as in (13), then we may encounter the following difficulties in using a Markov chain for obtaining an authority or a utility score for Web documents:

1. There are Web documents that do not contain any hyperlinks to other documents. In this case, condition (2) is not satisfied. Therefore, we cannot define a Markov chain from the probability transition matrix.

2. Even if the condition (2) is satisfied, all transient states have $\lim_{n\to\infty} p_{ij}^n = 0$, independently from the number of links that point to these states. Therefore this limit cannot be used as a score, since only persistent states would have a significant prestige (or quality) score.

There are two possible ways to overcome the above two problems:

1. We link all states by assigning a new probability $p_{ij}^* \neq 0$ in a suitable way, such that $|p_{ij}^* - p_{ij}| < \epsilon$. In this way all states become persistent. In other words the Web graph is transformed into a single irreducible closed set, namely the set of all states. Therefore, all states receive a positive prestige score. This approach is used in PageRank, where the assumed random surfer may randomly jump with a finite probability to any Web document.

2. We extend the original graph G to a new graph G^*. The new states of the extended graph G^* are all and only all the persistent states of the graph G^*. The scores of all states in the original graph, whether transient or persistent, will be uniquely associated to the scores of these persistent states in the new graph.

In the following section, we explore the second approach in order to overcome the above mentioned problems and define the Absorbing Model.

2.4 Definition of the Absorbing Model

The Absorbing Model is based on a simple transformation of the Web graph. 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 $\{C_i, \ldots, C_n\}$ 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 direct access only to its corresponding clone, but not to other clones. Since the clones are absorbing states, they do not have direct access to any state except to themselves. The Absorbing Model is formally introduced as follows:

Definition 1 *Let G = (D,R) be the graph consisting of the set D of 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 $d_{N+i}, i = 1, \ldots, N$, 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:*

$$R(d_i^*, d) = R(d, d_i^*) = 0, d \neq d_i^*, i = 1, \ldots, N \text{ except for:}$$
$$R(d_i, d_i^*) = 1$$
$$R(d_i^*, d_i^*) = 1$$

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

$$p_{ij} = \frac{R(d_i, d_j)}{|\{d_j : R(d_i, d_j) = 1\}|} \tag{14}$$

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

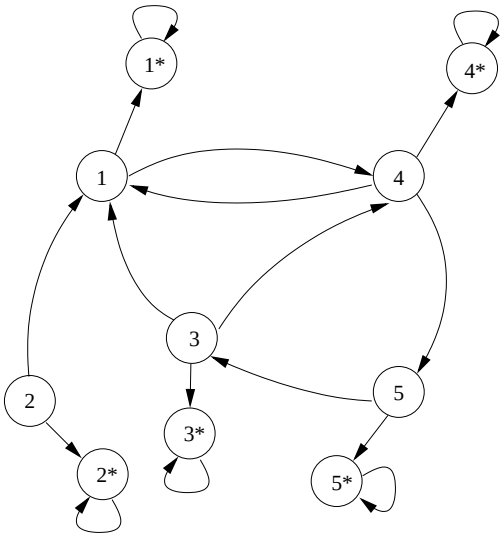

Fig. 2. The extended Markov Chain including the clone states.

Before continuing, we will give an example that illustrates the transformation of the graph. In Figure 1, we have shown a graph that represents a part of the Web. Figure 2 shows the same graph, transformed according to the definition of the 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^*. The introduced transformation results in removing any absorbing states from the original Web graph, as there are no closed sets consisting of any of the original states.

Hence, with the introduction of the clone nodes, all the original states $d_j, j = 1, \ldots, N$ become transient, while all the clone states $d_j^*, j = 1, \ldots, N$ are the only persistent states. In

other words, for the states in the original Markov chain we have:

$$p_{jk}^n \to 0,\ k = 1, \ldots, N \tag{15}$$

while for the clone states we have:

$$p_{jk}^n \to u_{jk},\ k = N+1, \ldots, 2N \tag{16}$$

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 $s(d_k)$ of a state d_k to be given by the unconditional probability of reaching its clone state d_k^*:

$$s(d_k) = \sum_j p_j u_{jk^*} \tag{17}$$

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

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.

At this stage, we would like to point out 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 to the model. For example, the use of the content retrieval scores as the prior probabilities results in a simple and principled way to combine dynamically content and link analysis [7], similarly to the extensions of HITS [9]. 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. This flexibility of the Absorbing Model enables its application in either a query-dependent, or a query-independent way.

In this paper, we focus on defining the prior probabilities independently of the content retrieval, and introduce the authority-oriented (Section 3) and the utility-oriented (Section 4) interpretations of the Static Absorbing Model.

3 The Static Absorbing Model

We introduce the Static Absorbing Model for measuring authority in a query-independent way. We define the model in Section 3.1 and evaluate it, along with PageRank, for TREC11 and TREC12 topic distillation tasks in Section 3.2.

3.1 Definition of the Static Absorbing Model

From the possible ways to define the prior probabilities independently of the queries, such as the document's length, or its URL type [20], one option 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:*

$$p_k = \frac{1}{2N} \tag{18}$$

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 (17) and (18) as follows:

$$s(d_j) = \sum_i p_i u_{ij^*} = \sum_i \frac{1}{2N} u_{ij^*} \propto \sum_i u_{ij^*} \tag{19}$$

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:

$$U = C^a \cdot L^b,\ a + b = 2 \tag{20}$$

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:

$$U_i = s(d_i|q) \cdot s(d_i) \tag{21}$$

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

3.2 Evaluation of the Static Absorbing Model

To test the effectiveness of the Static Absorbing Model, we have performed experiments using a standard Web test collection, namely the .GOV, which was used for the Web tracks of TREC11 [17], TREC12 [18] and TREC13 [21].

3.2.1 Experimental Setting

The .GOV collection is a standard TREC Web test collection, consisting of approximately 1.25 million Web documents. During indexing, stopwords were removed, and Porter's stemming algorithm was applied.

We employed the queries and relevance assessments from the topic distillation tasks of TREC11 and TREC12. Both tasks involve finding useful entry points to sites that are relevant to the query topics. However, a difference between the two tasks is that the relevant documents for the TREC12 topics were restricted to be homepages of relevant sites. This resulted in a lower number of relevant documents, less than 10 relevant documents for many queries, and thus it would not be theoretically possible to obtain 100% precision at 10 documents, which was the evaluation measure for TREC11. For this reason, the TREC Web track organisers chose the R-Precision (precision after R documents have been retrieved, where R is the number of relevant documents for the query) as the official evaluation measure [18]. We will use average precision and precision at 5 and at 10 documents for both TREC11 and TREC12. In addition, we will report R-Precision for the TREC12 experiments.

For the content analysis, we employed three different and independent weighting schemes. The first is the well-established $BM25$ [22], where we empirically set $b = 0.72$, $k_1 = 1$ and $k_3 = 1000$. The two other weighting schemes are $I(n_e)C2$ and $PL2$, from Amati and Van Rijsbergen's Divergence from Randomness (DFR) probabilistic framework [23]. For these weighting schemes, the weight of a query term within a document is given by the respective formulae:

$$\begin{aligned} weight_{PL2}(t) &= \Big(tfn_1 \cdot \log_2 \frac{tfn_1}{\lambda} + \\ &\quad + (\lambda + \frac{1}{12 \cdot tfn_1} - tfn_1) \cdot \log_2 e \\ &\quad + 0.5 \cdot \log_2 (2\pi \cdot tfn_1) \Big) \cdot \frac{1}{tfn_1 + 1} \end{aligned}$$

$$\begin{aligned} weight_{I(n_e)C2}(t) &= \frac{F+1}{\text{doc\_freq} \cdot (tfn_2 + 1)} \cdot \\ &\quad \cdot \left(tfn_2 \cdot \ln \frac{N+1}{n_e + 0.5} \right) \end{aligned}$$

where:

- $tfn_1 = \text{term\_freq} \cdot \log_2 \left(1 + c \cdot \frac{\text{average\_document\_length}}{document\_length} \right)$,
- $tfn_2 = \text{term\_freq} \cdot \ln \left(1 + c \cdot \frac{\text{average\_document\_length}}{document\_length} \right)$,

- N is the size of the collection,
- F is the within-collection term-frequency,
- $n_e = N \cdot \left(1 - \left(\frac{1}{N}\right)^{\text{Freq}(t|\text{Collection})}\right)$,
- λ is the mean and variance of the assumed Poisson distribution for the within-document term frequency. It is equal to $\frac{F}{N}$, where $F \ll N$,
- term_freq is the within-document term-frequency,
- doc_freq is the document-frequency of the term.

The weight of a document d for a query q is given by:

$$weight(d,q) = \sum_{t \in q} qtf \cdot weight_x(t) \tag{22}$$

where $weight_x(t)$ is the weight of a document for a query term t, as defined above for the weighting schemes $PL2$ and $I(n_e)C2$, and qtf is the frequency of the query term t in the query q. The only parameter of the DFR framework is set equal to $c = 1.28$ automatically, according to a method proposed by He and Ounis for the tuning of term frequency normalisation parameters [24]. For our experiments, we used Terrier [25], an information retrieval platform for large-scale experimentation, which provides a range of DFR and classic retrieval models.

In order to combine hyperlink analysis with effective content retrieval approaches, we use content-only retrieval for TREC11, as it was the most effective approach [17]. For the TREC12 experiments, we extend the documents with the anchor text of their incoming links and use this representation as a baseline. This retrieval approach outperforms content-only retrieval significantly [18].

For the hyperlink analysis, the Static 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. Note that the computational overhead due to the introduction of the clone states was insignificant. In addition, we ran experiments where we use PageRank, instead of the Absorbing Model in the Cobb-Douglas utility function (Equation 20)), with $a = b = 1$.

3.2.2 Evaluation

The evaluation of both the authority-oriented Static Absorbing Model (SAM) and PageRank (PR), combined with the different weighting schemes using the Cobb-Douglas utility function, is shown in Table 1 for TREC11 and Table 2 for TREC12 respectively. The indices of SAM and PR in the tables denote the weighting method used for the content analysis. The best official run submitted to TREC11 topic distillation task achieved 0.2510 precision at 10, while the highest precision at 10 for TREC12 was 0.1280 and the highest R-Precision was

Table 1. Authority-oriented experiments with Static Absorbing Model and PageRank for the TREC11 topic distillation topics. The content analysis is based on the textual content of documents. Prec@x stands for precision at x documents.

	Av. Prec.	Prec@5	Prec@10
$BM25$	0.1919	0.2939	0.2408
SAM_{BM25}	0.0022	0.0082	0.0041
PR_{BM25}	0.0034	0.0041	0.0204
$PL2$	0.2058	0.3102	0.2694
SAM_{PL2}	0.0028	0.0082	0.0041
PR_{PL2}	0.0039	0.0163	0.0265
$I(n_e)C2$	0.1983	0.3061	0.2490
$\text{SAM}_{I(n_e)C2}$	0.0024	0.0082	0.0041
$\text{PR}_{I(n_e)C2}$	0.0037	0.0082	0.0245

Table 2. Authority-oriented experiments with Static Absorbing Model and PageRank for the TREC12 topic distillation topics. The content analysis is based on the textual content and anchor text of documents. Prec@x stands for precision at x documents.

	Av. Prec.	Prec@5	Prec@10	R-Prec.
$BM25$	0.1212	0.1280	0.1020	0.1293
SAM_{BM25}	0.0063	0.0040	0.0040	0.0113
PR_{BM25}	0.0049	0.0080	0.0060	0.0030
$PL2$	0.1273	0.1240	0.1020	0.1325
SAM_{PL2}	0.0074	0.0080	0.0040	0.0117
PR_{PL2}	0.0076	0.0080	0.0100	0.0123
$I(n_e)C2$	0.1195	0.1240	0.0940	0.1222
$\text{SAM}_{I(n_e)C2}$	0.0063	0.0040	0.0040	0.0113
$\text{PR}_{I(n_e)C2}$	0.0054	0.0080	0.0060	0.0101

0.1636 (these figures correspond to two different runs). The results show that for both SAM and PageRank, the authority-oriented approach is not effective for retrieval on the specific collection, independently of the weighting function used. In the remainder of this section, we look into possible explanations of this result.

First, we consider the effect of combining scores from different score distributions. Aslam and Montague propose three conditions, which must be satisfied for successfully combining evidence [26]. One of these conditions is that the combined score distributions should be on the same scale. However, this condition does not hold when we combine content and hyperlink analysis. Manmatha et al. [27] model the score distribution of the retrieved documents as a mixture of two distributions: a Gaussian distribution for the scores of the relevant documents, and an exponential distribution for the scores of the non-relevant documents. On the other hand, Pandurangan et al. [28] suggest that the values of PageRank follow a power law. Similarly, we have found that the probability distribution of the Absorbing Model scores follow a power law with exponent -1.34. In order to smooth the difference between the distributions of content analysis and hyperlink analysis scores, we choose to modify the hyperlink analysis scores, because the corresponding power law is highly skewed. We update Equation (21) by smoothing the hyperlink analysis scores with a logarithmic function, as follows:

$$U_i = s(d_i|q) \cdot \log_2(shift \cdot s(d_i)) \tag{23}$$

where $shift$ is a parameter introduced to (i) ensure that we obtain only positive values and (ii) adjust the influence of the authority-oriented hyperlink analysis score. When $shift$ is low, the hyperlink analysis scores modify the content-based ranking of documents significantly.

On the other hand, for the higher values of $shift$, the hyperlink analysis scores do not alter significantly the content-based ranking. The parameter $shift$ is set to 10^k, where k is an integer in the range $[4, 12]$. We have found that $s(d_i) > 10^{-4}$, so the lowest value 10^4 ensures that the logarithm is always positive. We select the upper limit for the parameter $shift$ to be equal to 10^{12}, in order not to over-smooth the scores $s(d_i)$.

Figures 3 and 4 contain the results from the experiments, where we adjust the value of $shift$ and consequently, the smoothing of the hyperlink analysis scores. From these figures, we can see that the effectiveness of the authority-oriented hyperlink analysis is still less than that of the content-only baselines, for both TREC tasks. In addition, we can see that when $shift = 10^4$, that is when the influence of hyperlink analysis is higher, PageRank is more effective than the Absorbing Model. However, the Absorbing Model combines more effectively with content analysis, as the value of $shift$ increases and the hyperlink analysis scores are smoothed. For the higher values of $shift$, the retrieval effectiveness approaches that of the baselines. Thus, smoothing the hyperlink analysis scores, and making them comparable to the content analysis scores, has a positive effect on the retrieval effectiveness. This confirms the appropriateness of the compatibility condition between the scores of different systems or retrieval approaches, and it is a point we will investigate further in future work.

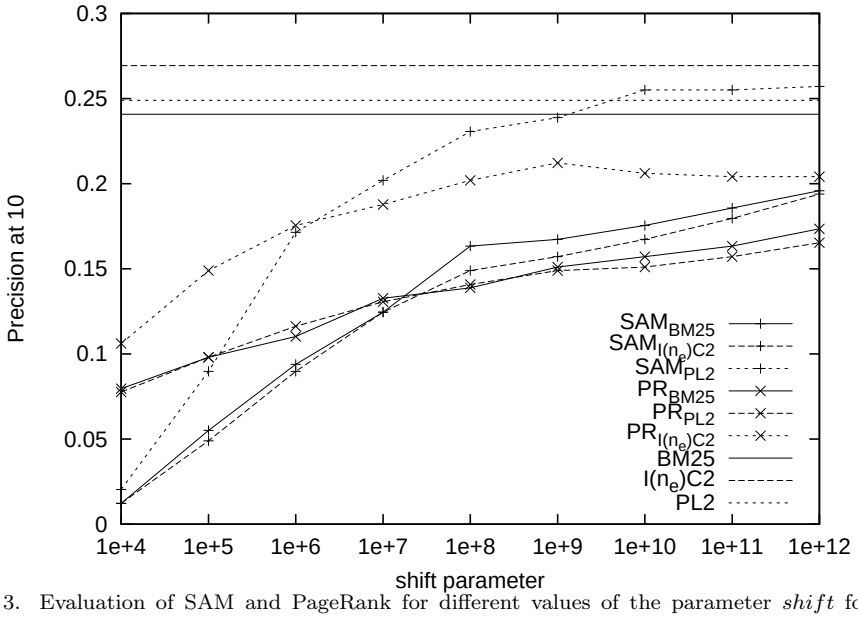

Fig. 3. Evaluation of SAM and PageRank for different values of the parameter $shift$ for the TREC11 topic distillation task. The content analysis is based on the textual content of documents.

Moreover, a close look at the collection and the topic distillation tasks suggests that this authority-oriented approach may not be suitable for application on a collection, where all the resources 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, it cannot be used to leverage authority. Even though searching the .GOV collection is more similar to searching a *small Web* [29], there are no

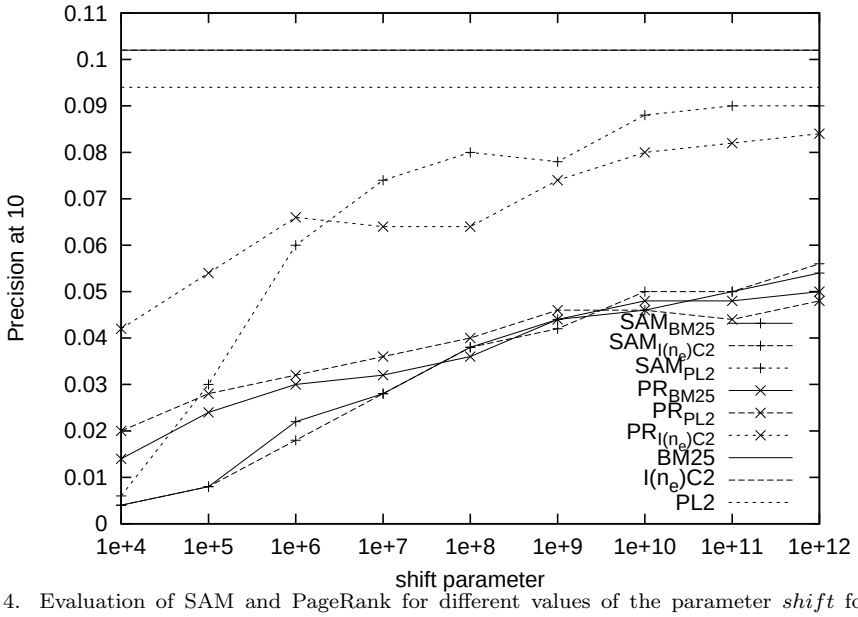

Fig. 4. Evaluation of SAM and PageRank for different values of the parameter $shift$ for the TREC12 topic distillation task. The content analysis is based on the textual content of documents and the anchor text of their incoming hyperlinks.

other public large-scale Web test collection with relevance judgements.

In addition, it could be the case that the 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 [30]. The impact factor is not computed for single papers, but for aggregates of papers, which are published in the same journal. However, it is not straightforward to relate the fields of citation and hyperlink analysis, since the motivations for adding citations in a scientific paper are different from the motivations for adding a hyperlink to a Web document [31]. In the context of a query-biased retrieval methodology for topic distillation, we have found that aggregates of Web documents provide us with useful information regarding the most appropriate retrieval approaches [32].

4 The Static Utility Absorbing Model

In this section, we focus on a different approach to hyperlink analysis, where we consider the documents' utility. We define the utility-oriented instance of the Absorbing Model in Section 4.1 and evaluate it in Section 4.2.

4.1 Definition of the Static Utility Absorbing Model

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 this paper, we focus on a query-independent hyperlink analysis model and the concept of utility is based on the number of outgoing links from a document. In the context of the Web, this simplified concept of utility could be manipulated by the authors of documents. However, we believe that it may be more appropriate for the .GOV collection, and other controlled environments. We have also investigated the concept of utility of documents, based on a combination of content and hyperlink analysis. In the context of a decision mechanism for selecting appropriate retrieval approaches on a per-query basis, this approach has led to important improvements in retrieval effectiveness [33].

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 $-log_2(s(d_j))$ [34]. As already mentioned in Section 2, the Absorbing Model score of a document depends on both the 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.

For the combination of evidence, we employ again the Cobb-Douglas utility function, (Equation (20)), with exponents $a = b = 1$. Differently from SAM, we replace the Absorbing Model score with its informative content:

$$U_i = s(d_i|q) \cdot (-log(s(d_i)) \tag{24}$$

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

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.

4.2 Evaluation of the Static Utility Absorbing Model

For testing the effectiveness of SUAM, we experiment in the setting described in Section 3.2.1. As we can see from the results presented in Tables 1 and 2 for SAM and Tables 3 and 4 for SUAM, the utility-oriented Absorbing Model is considerably better than SAM. Both average precision and precision at 10 are at the levels of the content-only baseline for TREC11, and they are better for all weighting schemes in TREC12. In addition, in our TREC11 experiments, precision at 5 increases for $I(n_e)C2$ and remains stable for BM25 and PL2. For TREC12, when we combine SUAM with any of the three weighting schemes we test, precision at 5 and precision at 10 increase. Moreover, R-Precision increases for the TREC12 experiments only when SUAM is combined with $PL2$.

Overall, the utility-oriented hyperlink analysis improves the retrieval effectiveness, particularly for TREC12, where relevant documents are restricted to the homepages of sites. Indeed, we expect that the relevant documents will have more outgoing links, in order to facilitate the navigation of users, and consequently, they will get a higher SUAM score. The stability of the retrieval effectiveness for TREC11 also shows that SUAM is a robust model, even when the relevant documents are not restricted to the homepages of sites [17].

Table 3. Static Utility Absorbing Model results for TREC11 topic distillation. The content analysis is based on the textual content of documents.

	Average Precision	Precision at 5	Precision at 10
$BM25$	0.1919	0.2939	0.2408
SUAM_{BM25}	0.1861	0.2898	0.2306
$PL2$	0.2058	0.3102	0.2694
SUAM_{PL2}	0.2034	0.3102	0.2510
$I(n_e)C2$	0.1983	0.3061	0.2490
$\text{SUAM}_{I(n_e)C2}$	0.1906	0.3224	0.2306

Table 4. Static Utility Absorbing Model results for TREC12 topic distillation. The content analysis is based on the textual content and anchor text of documents.

	Average Precision	Precision at 5	Precision at 10	R-Precision
$BM25$	0.1212	0.1280	0.1020	0.1293
SUAM_{BM25}	0.1247	0.1600	0.1200	0.1232
$PL2$	0.1273	0.1240	0.1020	0.1325
SUAM_{PL2}	0.1357	0.1360	0.1060	0.1401
$I(n_e)C2$	0.1195	0.1240	0.0940	0.1222
$\text{SUAM}_{I(n_e)C2}$	0.1240	0.1360	0.1200	0.1165

For the remainder of this section, we will perform a detailed analysis of the results. We will use the weighting scheme $I(n_e)C2$, which resulted in improvements for both TREC11 and TREC12 topic distillation tasks. In Tables 5 and 6, we compare the effectiveness of $I(n_e)C2$ and $\text{SUAM}_{I(n_e)C2}$ in TREC11 and TREC12 respectively. The first column refers to the evaluation measure used for comparing the two approaches. The next three columns correspond to the number of queries for which we observed an improvement (+), a loss in precision (-), or where the effectiveness remained the same (=). The last column presents the resulting p values from the Wilcoxon's signed rank test for paired samples, which shows that $\text{SUAM}_{I(n_e)C2}$ resulted in a significant improvement over $I(n_e)C2$, with respect to precision at 10 documents for the TREC12 topic distillation task. An interesting difference between the two topic distillation tasks is that the improvements in precision at 5 and precision at 10 are not consistent. Precision at 5 documents increases for more queries in TREC11 than in TREC12. On the other hand, precision at 10 benefits more for TREC12. We believe that this difference is a result of the lower number of relevant documents found for the TREC12 queries, which results in smaller improvements in precision at 5.

Overall, the application of SUAM results in increased precision amongst the top ranked documents for both TREC tasks. The obtained results indicate that the utility-oriented link analysis is more appropriate for the topic distillation tasks 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 robust than the latter, in both TREC11 and TREC12 topic distillation tasks.

Table 5. Query-by-query analysis of $I(n_e)C2$ versus $\text{SUAM}_{I(n_e)C2}$ for TREC11 topic distillation. The content analysis is based on the textual content of documents.

Measure	+	-	=	p (Singed ranks test)
Average Precision	21	27	1	0.154
Precision at 5	10	7	32	0.515
Precision at 10	7	13	29	0.207

Table 6. Query-by-query analysis of $I(n_e)C2$ versus $\text{SUAM}_{I(n_e)C2}$ for TREC12 topic distillation. The content analysis is based on the textual content and anchor text of documents.

Measure	+	-	=	p (Singed ranks test)
Average Precision	26	22	2	0.426
Precision at 5	7	6	37	0.479
Precision at 10	14	4	32	0.015
R-Precision	11	6	33	0.495

5 Extended experiment with the Static Utility Absorbing Model

In order to further examine the Static Utility Absorbing Model, we investigated the effect of adjusting the parameters a and b in the utility function (20). The exponents represent the relative importance of each of the components used in the combination of evidence. Note that we effectively introduce only one parameter in the model, because the sum of the exponents should be constant, i.e. equal to 2. Similarly to the detailed analysis of the previous section, we will use $\text{SUAM}_{I(n_e)C2}$ since it resulted in improvements for both TREC tasks.

We have conducted an extensive experiment in which we set the exponents a and b to values between 0 and 2 in steps of 0.1. In Figures 5 and 6, the evaluation output of $\text{SUAM}_{I(n_e)C2}$ is presented for the different values of the exponent b. The exponent for the content-based module is $a = 2 - b$, and the precision value of the $I(n_e)C2$ content-only baseline corresponds to the points for $b = 0$. We can see that $\text{SUAM}_{I(n_e)C2}$ is relatively stable across a wide range of values of b for TREC11 (see Figure 5) and results in improved retrieval effectiveness for TREC12 (see Figure 6). Its performance decreases rapidly for both TREC tasks when b approaches 2.0.

More specifically for TREC11, the highest precision at 5 is 0.3306, which is obtained for $b = 0.6$. Both precision at 10 and average precision remain stable for a wide range of b values. For TREC12, the improvements over the content-only baseline are more evident. In Figure 6, we can see that there are improvements for all reported evaluation measures. In addition, the bold points in the figure correspond to points, where we found that $\text{SUAM}_{I(n_e)C2}$ improved significantly the corresponding measure, according to the Wilcoxon's signed rank test ($p \leq 0.038$ for precision at 10 and $p \leq 0.018$ for average precision). The highest precision at 10 is 0.1240 and the highest average precision is 0.1348.

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. 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 for the TREC11 experiments, the first group consists of 22 queries (for 5 out of the 22 queries in this group, no relevant documents were retrieved by our system) and the second one consists of 27 queries. For TREC12, there are 30 queries in the first group (for 2 out of the 30 queries, no relevant documents were retrieved) and 20 queries in the second one.

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-

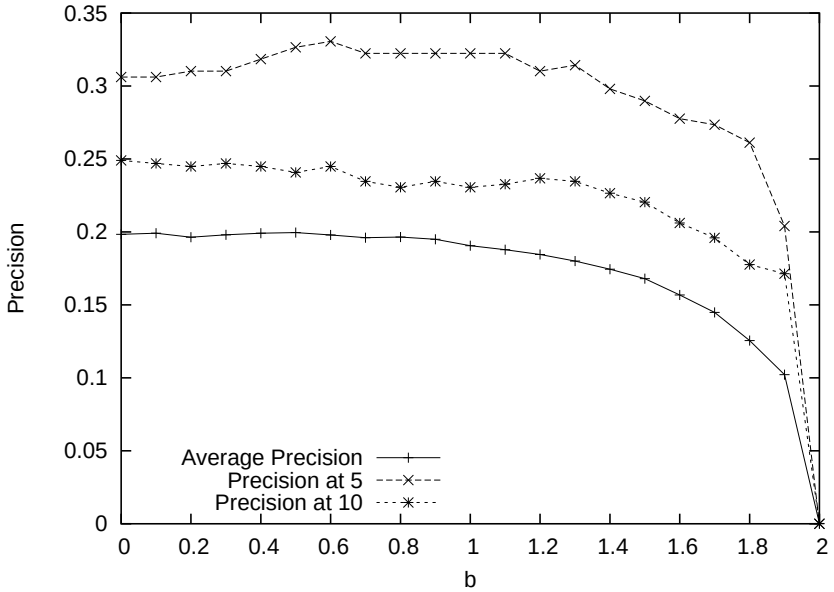

Fig. 5. Precision for different values of the exponents, for TREC11 topic distillation. The $I(n_e)C2$ baseline corresponds to the point, where $b = 0$.

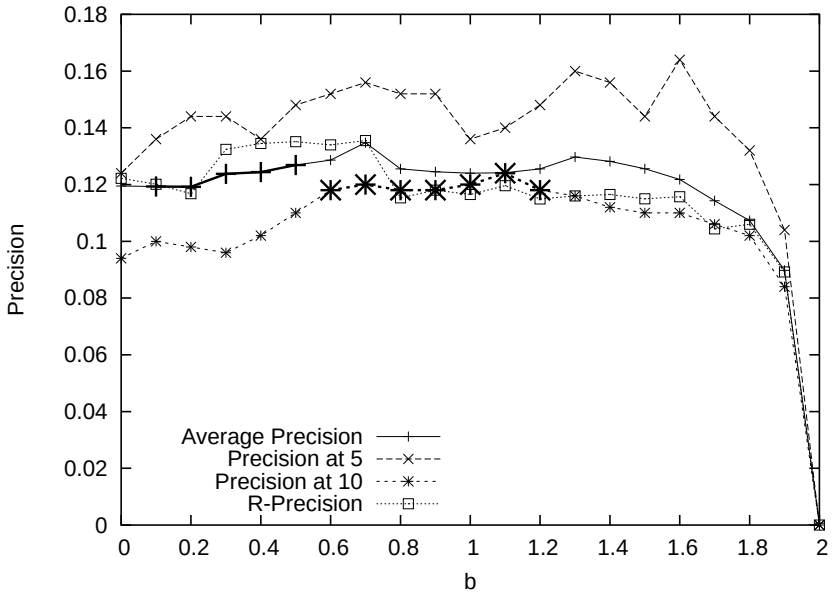

Fig. 6. Precision for different values of the exponents, for TREC12 topic distillation. The $I(n_e)C2$ baseline corresponds to the point, where $b = 0$.

query basis for the corresponding measure of retrieval effectiveness. Tables 7 and 8 summarise 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, the retrieval effectiveness would improve significantly for both TREC11 and TREC12 topic distillation tasks (see the row Maximum1

Table 7. Comparison between $I(n_e)C2$ and the ideal cases for TREC11 topic distillation.

	Average Precision	Precision at 5	Precision at 10
$I(n_e)C2$	0.1983	0.3061	0.2490
Maximum1	0.2195 ($p < 5 \cdot 10^{-4}$)	0.3592 ($p < 5 \cdot 10^{-4}$)	0.2714 ($p = 0.038$)
Maximum3	0.2067 ($p = 0.029$)	0.3388 ($p = 0.100$)	0.2653 ($p = 0.096$)

Table 8. Comparison between $I(n_e)C2$ and the ideal cases for TREC12 topic distillation.

	Average Precision	Precision at 5	Precision at 10	R-Precision
$I(n_e)C2$	0.1195	0.1240	0.0940	0.1222
Maximum1	0.1781 ($p < 5 \cdot 10^{-4}$)	0.1840 ($p = 0.001$)	0.1320 ($p < 5 \cdot 10^{-4}$)	0.1608 ($p = 0.001$)
Maximum3	0.1378 ($p = 0.001$)	0.1680 ($p = 0.028$)	0.1280 ($p = 0.001$)	0.1305 ($p = 0.096$)

from Tables 7 and 8, respectively).

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 correspond to just the third best average precision per query, precision amongst the top ranked documents would still improve considerably (see Maximum3 in Tables 7 and 8 for TREC11 and TREC12, respectively). More specifically, using the Wilcoxon's signed ranks test, we can see that we would obtain significant improvements in average precision and precision at 5 documents, for both TREC11 and TREC12 tasks. In addition, precision at 10 documents would increase significantly for the TREC12 topic distillation task.

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 $I(n_e)C2$. For example, if we maximised the average precision for TREC11, then precision at 10 documents would be 0.2714, while if we aimed at maximising precision at 10, then the obtained precision would be 0.2959. In the same way, maximising average precision for TREC12 results in 0.1320 precision at 10 documents, while if we chose to maximise precision at 10 documents, we would get a maximum of 0.1460. The values of the parameters a and b that maximise the average precision result in maximum precision at 10 documents for 39 queries from TREC11 and for 40 queries from TREC12. These results show the high correlation between the average precision and the precision at 5 and 10 documents.

The usefulness of the obtained results lies in the fact that we have a guideline for the optimal combination of content and hyperlink analysis with SUAM. We can employ this guideline to evaluate the effectiveness of methods for setting the parameters a and b to appropriate values automatically, similarly to our work in [33].

6 Conclusions

In this paper, we have presented the Absorbing Model, a hyperlink analysis model for Web information retrieval, based on Markov chains. Differently from PageRank, the Absorbing Model is based on introducing an absorbing clone node for each node in the original graph, so that in the extended graph, all original nodes become transient and only the clone nodes are persistent. The Absorbing Model does not possess a stationary probability distribution.

Therefore, its scores depend on the prior probabilities and it can be applied either during indexing, or during query-time. This allows for a natural combination of evidence from content and hyperlink analysis, in a static (query-independent), or a dynamic (query-dependent) way. In this paper, we focus on using the Absorbing Model in a static way. In the future, we aim to evaluate the dynamic version of the Absorbing Model, where the prior probabilities of documents will correspond to the content retrieval scores [7].

We have proposed two different interpretations of the Absorbing Model, 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. For both models, the combination of evidence from the content and hyperlink analysis is achieved by employing the Cobb-Douglas utility function.

We have performed experiments with both approaches, using the .GOV Web test collection and the topic distillation queries from TREC11 and TREC12. We have found that, although smoothing the hyperlink analysis scores benefits the retrieval effectiveness, the authority-oriented hyperlink analysis is not highly effective for this test collection. The .GOV collection contains documents from a controlled and high quality domain, which resembles more a small Web search environment [29]. On the other hand, the Static Utility Absorbing Model is stable and improves precision among the top ranked documents for the same collection in both TREC11 and TREC12. 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 used test collection.

In addition, we have shown that in the ideal case, where the best values for the parameter b of the Cobb-Douglas utility function were chosen automatically, the retrieval effectiveness would improve significantly. This result is important, because it shows the potential benefits from hyperlink structure analysis and query-biased retrieval in the context of TREC-like experiments. Indeed, we have employed a decision mechanism for selecting appropriate retrieval approaches on a per-query basis [32, 33], and obtained important improvements in retrieval effectiveness.

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).

References

1. 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.
2. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, Stanford, CA, 1998.
3. T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, pages 517–526. ACM Press, 2002.

4. 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.
5. S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the Block Structure of the Web for Computing PageRank. Technical report, Stanford University, Stanford, CA, 2003.
6. G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th International Conference on World Wide Web*, pages 271–279. ACM Press, 2003.
7. 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.
8. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
9. 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.
10. S. Chakrabarti, B.E. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*, pages 307–318. ACM Press, 1998.
11. 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.
12. D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proceedings of the 17th International Conference on Machine Learning*. ACM Press, 2000.
13. H.-Y. Kao, M.-S. Chen, S.-H. Lin, and J.-M. Ho. Entropy-based link analysis for mining web informative structures. In *Proceedings of the 11th International ACM Conference on Information and Knowledge Management (CIKM)*, pages 574–581. ACM Press, 2002.
14. P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura, and I. Silva. Local versus global link information in the web. *ACM Transactions on Information Systems*, 21:42–63, January 2003.
15. W.B. Croft and H. Turtle. A retrieval model incorporating hypertext links. In *Proceedings of the 2nd Annual ACM Conference on Hypertext*, pages 213–224. ACM Press, 1989.
16. 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.
17. 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.
18. N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC 2003 Web Track. In *NIST Special Publication 500-255: The Twelfth Text REtrieval Conference (TREC 2003)*, pages 78–92, 2003.
19. W. Feller. *An Introduction to Probability Theory and its Applications, volume 1, 2nd edition*. John Wiley and Sons, 1957.
20. 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.
21. N. Craswell, and D. Hawking. Overview of the TREC 2004 Web Track. In *The Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.
22. S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241. Springer-Verlag New York, Inc., 1994.
23. G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.
24. 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.
25. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald and D. Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on Information Retrieval (ECIR05)*. Springer, 2005.
26. J. A. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 276–284. ACM Press, 2001.
27. R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–275. ACM Press, 2001.
28. G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.
29. G.-R. Xue, H.-J. Zeng, Z. Chen, W.-Y. Ma, H.-J. Zhang, and C.-J. Lu. Implicit link analysis for small web search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 56–63. ACM Press, 2003.
30. E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.
31. M. Thelwall. What is this link doing here? beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3), 2003.
32. V. Plachouras, I. Ounis, and F. Cacheda. Selective combination of evidence for topic distillation using document and aggregate-level information. In *Proceedings of RIAO 2004: Coupling approaches, Coupling Media and Coupling Languages for Information Retrieval*. C.I.D, 2004.
33. V. Plachouras and I. Ounis. Usefulness of hyperlink structure for query-biased topic distillation. In *Proceedings of the 27th annual international SIGIR Conference on Research and Development in Information Retrieval*, pages 448–455. ACM Press, 2004.
34. K. Popper. *The Logic of Scientific Discovery*. Hutchinson & Co., London, 1959.