The Dynamic Absorbing Model for the Web

Gianni Amati[*], Iadh Ounis, Vassilis Plachouras
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, U.K.

Abstract

In this paper we propose a new theoretical method for combining both content and link structure analysis for Web information retrieval. This approach is based on performing a random walk on a modified Markov chain, induced from the Web graph. This new model is applicable either in a static way, where a global prestige score is computed for every document, or in a dynamic way, where the link structure of the results from a first pass retrieval is taken into account. The results of experiments on the TREC WT10g and .GOV collections show that it is a robust method, particularly suitable for dynamic link analysis. Moreover, we show that it outperforms PageRank under the same experimental settings.

1 Introduction

Combining efficiently and in a principled way the evidence from both content and link structure analysis is a crucial issue in Web information retrieval. Most of the methodologies proposed so far are either based on ad-hoc heuristics, or need computationally intensive training for optimising their parameters.

One of the most popular algorithm for hyperlink analysis is PageRank [5]. PageRank computes the Web popularity of a document by an invariant distribution of a Markov chain, which is derived only from the structure of the Web graph. Therefore, PageRank assigns global prestige scores to Web documents independently from content. There are many extensions and variants of PageRank. For example Richardson and Domingos [24] and Haveliwala [12] try to relate popularity with content using a predefined set of popular

[*] Also affiliated to Fundazione Ugo Bordoni.

queries, or topics. These are attempts to overcome the independence of the prestige score from topic relevance, which instead is the *prior* probability of retrieving a document by content. Unfortunately, the invariant distributions of PageRank are called "invariant" just because they are not affected by the choice of the priors. Thus the combination of prestige scores and content scores must be obtained with a different approach.

Recognising the difficulties in combining both content and link analysis, we look into a possibly parameter-free approach to link analysis, in which the combination of evidence from content analysis and the structure of hyperlinks is natural and intuitive. We propose a new methodology, which we call the Absorbing Model. This approach is based on performing a random walk on a Markov chain induced from a modified Web graph. Although the invariant distribution for the Absorbing Model does not exist, we exploit this and the limiting properties of Markov chains to define the prestige score as a well founded combination of content and link analysis. This prestige score depends on both the link structure of the Web, as well as on the prior probabilities assigned to documents. We have two possible ways to define the prior probabilities. For the *static* application of the Absorbing Model, we can assign a uniform probability distribution and compute a prestige score during indexing. Alternatively, we can assign the normalised content scores, merging thus the content and link analysis dynamically query-by-query.

We get promising results by evaluating the effectiveness of the Absorbing Model with respect to the test collections used in TREC10 topic relevance task (or ad-hoc task) and TREC11 topic distillation and named page finding tasks. It has been observed in the last two TREC conferences [13, 8] that these collections do not favour the application of link analysis. However, our finding is controversial. While PageRank seems to be largely detrimental to the retrieval effectiveness in both collections, we observed a striking difference in the performance of the static application of the Absorbing Model over the two collections. Only when relevance is defined by the ad hoc task, and not by the topic distillation task, the Absorbing Model does not degrade with respect to the content only analysis. This seems to indicate that the hypothesis on uniform distribution is not adequate for the topic distillation task. The indication is confirmed by the dynamic application of the Absorbing Model only to the subgraph of a first pass top ranked documents, a more stable and effective approach that outperforms content-only retrieval.

Finally, we evaluate the effectiveness of incorporating the anchor text of links where we found that anchor text is effective only in the named page finding task.

The rest of this paper is organised as follows. In Section 2 we review related work in the research area of link analysis and we introduce the basic properties of Markov chains. The definition of the Absorbing Model follows in Section 3 and in Section 4 we present the experiments we performed and the corresponding results. Section 5 closes with the conclusions we have drawn from this work and some points of interest which require further examination.

2 Related work

In this section, we provide a brief overview of the related work in the field of link analysis for Web Information Retrieval (Section 2.1) and a formal description of Markov chains and their basic properties that we will use for defining the Absorbing Model (Sections 2.2, 2.3 and 2.4).

2.1 Link analysis for Web Information Retrieval

The Web is not the first context in which the analysis of interconnected documents has been applied. Earlier, the analysis of the references of scientific publications has been employed to estimate the impact factor of papers, or journals [11]. Other useful measures resulting from analysis of citations are the bibliographic coupling and the co-citation coupling [22]. Similar techniques have been proposed in the context of social sciences, where, for example, Milgram [20] analysed the connections between people.

The rapid growth of the Web, together with the need to find useful resources resulted in the requirement to detect quality resources, by estimating the authority of Web documents. The hyperlink structure provides the evidence for detecting the most authoritative documents, by exploiting the latent knowledge put by the authors of documents when they add hyperlinks to other documents. Two of the first methods proposed for link analysis on the Web are Kleinberg's HITS algorithm [17] and Brin and Page's PageRank [5].

Kleinberg in HITS suggests that Web documents can be classified into two categories, namely the hubs and the authorities, which ideally form bipartite graphs. The hubs are Web documents which point to many useful authority documents on a specific topic, while the authorities are Web documents that are pointed by many hubs on the same topic. Moreover, there is a mutual reinforcement relation between authorities and hubs: good authorities are pointed by good hubs and good hubs point to many good authorities. The algorithm works as follows: initially a set of Web documents is retrieved by using a standard search engine and this set is extended by adding documents that are pointed, or that point to the documents of the initial set. The adjacency matrix A of the graph that corresponds to the extended set is created and the principal eigenvectors of the matrices AA^T and A^TA are computed. The component of each document in the principal eigenvector of AA^T corresponds to its hub weight, while its component in the principal eigenvector of A^TA corresponds to its authority value.

The second algorithm, namely PageRank, proposed by Brin and Page [5], is based on the calculation of the invariant probability distribution of the Markov chain induced from the Web graph and returns a global authority score for each indexed Web document. The authority score of a document depends on the authority scores of the Web documents that point to it, and it is calculated by an iterative process. However, the corresponding Markov chain does not always guarantee the convergence of the process. For that, the structure of the Web is altered in the following way. A link is added from every Web document to each other. This transformation of the Web graph can be interpreted in the following way: a random user that navigates in the Web has two possibilities at each time; either to follow a link from the document that he is currently browsing, or to jump to a randomly selected Web document and continue his navigation from that document. The addition of this element of randomness results to a more stable algorithm [21] and guarantees the existence of the invariant distribution of the corresponding Markov chain.

Recent proposals for link analysis methods follow a probabilistic approach to the analysis of the Web documents structure. Cohn and Chang propose a method called PHITS, which is mathematically equivalent to Hofmann's Probabilistic Latent Semantic Indexing [14], for calculating authority and hub scores for Web documents. Moreover, Lempel and Moran introduce SALSA [18], a refinement of the HITS algorithm, where instead of analysing one random walk, there are two Markov chains, one for the hubs and one for the authorities. The random walk is performed by making at each point a step forwards immediately followed by a step backwards in the Markov chains. Additional refinements to the HITS algorithm are introduced by Borodin et al. [4].

Apart from the pure link analysis methods described above, there have been efforts for the effective combination of content and link-based evidence. Bharat and Henzinger refine HITS by incorporating the associated anchor text of the links in order to weight each link [3]. Moreover, Chakrabarti and Chakrabarti et al. use the DOM tree representation of Web documents to identify specific segments of documents about a single topic, which they call microhubs [6, 7]. A modified version of PageRank is proposed by Domingos and Richardson, where the links are weighted according to their similarity to the query, resulting to a query dependent authority measure [24]. Another extension to PageRank is introduced by Haveliwala for estimating a personalised authority measure for documents [12]. Instead of calculating a single global PageRank score for each document, different PageRank biased on specific topics are calculated and the final authority measure of documents is the combination of those PageRank scores.

Other methods, which do not extend directly HITS or PageRank, have been proposed for the combination of content and link-based evidence. Silva et al. employ Bayesian networks model and combine evidence from the content and the link structure of documents [26]. Savoy and Picard use a spreading activation mechanism where a fraction of the retrieval status values propagate through links [25], and Jin and Dumais propose a similar approach taking into account the similarity of documents, the similarity of the anchor text of links to the query and the prestige scores of documents [15].

2.2 Markov Chains

Since the Absorbing Model is a link analysis method that models the Web as a Markov chain, we introduce the concept of Markov chains and their basic properties. The notation used in this section and in the rest of the paper is similar to that of Feller [10].

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 to be selected 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 maybe 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 and the transition probabilities as a row-by-column matrix, 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 is said *closed* if any state inside the set can reach all and only all other states inside 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 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 anymore. A state which is not in any closed set is called *transient*. A transient state must reach at least one state in a closed set and 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 $C_1, C_2 \ldots$ 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*.

For a better understanding, we give an example of the above definitions. In Figure 1, the directed graph may be interpreted as the Markov Chain corresponding to a few Web documents with the arcs representing the links between documents and consequently the transitions between the 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 not 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.

2.3 Classification of states

The probability of reaching the state d_j from any initial state by any random walk $w = (d_i, \ldots, d_j)$ is thus, according to (6):

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

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}. Therefore, with these new transition probabilities, p'^n_{ii} will be greater than 0 and the periodic states d_j become aperiodic. Hence, we may assume that all states in a Markov chain are aperiodic [10].

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 any 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. This property makes an irreducible Markov chains to possess an invariant distribution, that is a distribution u_k such that

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

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:

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

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.

2.4 Modelling the hyperlinks of the Web

In this section we present formally 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 a prestige 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 $\lim\limits_{n\to\infty} p_{ij}^n = 0$, independently from the number of links that point to these states. Therefore this limit cannot be used as a prestige score, since only persistent states would have a significant prestige score.

There are two possible ways to overcome the two previous 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 relative to all states of the original graph, whether transient or persistent, will be uniquely associated to the scores of these persistent states of 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.

3 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 nodes, that is a set of closed sets containing only one element. The states $C_i, \ldots, C_n$ are called the *clones* of G and they are uniquely associated to the states of the graph G. Any state in G has access to its clone but not to other clones. Since the clones are absorbing nodes 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 representing the hyperlinks between documents. The graph G is extended by introducing N additional nodes $d_{N+i}, i = 1, \ldots, N$, called the clone nodes. These additional nodes are denoted as: $d_{N+i} = d_i^*$. The accessibility relation R is extended in the following way:

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

In Figure 1, a graph representing five Web documents is shown. According to the Absorbing Model's definition, this graph is transformed into the graph of Figure 2.

Hence, with the introduction of the clones, all the original states become transient, while the only closed sets are the singleton sets that contain each clone d_k^*. Moreover, one of the properties of the Absorbing Model is that the self-loops in the set of transient nodes do not influence the limits of the matrix M^n. The transition probabilities in the original graph are updated accordingly:

$$p_{ij} = \frac{R(i,j)}{o(i) + 1} \tag{14}$$

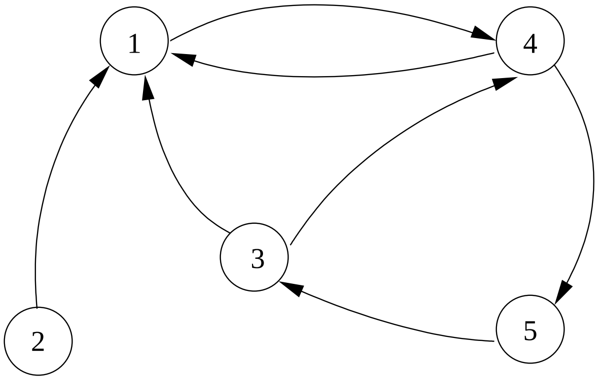

Figure 1: The Markov Chain representing the web graph

Then, for the nodes of the original graph we have:

$$p_{jk}^{n} \rightarrow 0 \quad (k = 1, \ldots, N) \tag{15}$$

whilst for the absorbing states we have:

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

where u_{jk} stands for the probability that a random walk starting from d_j passes through d_k.

The *score* of a state d_k is then given by the unconditional probability of reaching its clone state d_k^*, that is:

$$\sum_j p_j u_{jk^*} \tag{17}$$

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

While in PageRank the rows in the limiting matrix are equal, in the Absorbing Model they are in general different, that is $u_{ik} \neq u_{jk}$ for different i, j, so that the limiting matrix does not correspond to an invariant distribution. Hence, the row vector P of priors can significantly modify the score of a document. We therefore define the prior probabilities P into two different ways:

Definition 2 Static mode priors: we assume that the prior probability of selecting document d_k is uniformly distributed:

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

We should note that the number $2N$ refers to the total number of states of the new graph G^*, that is the total number of documents plus the total number of their corresponding clone states.

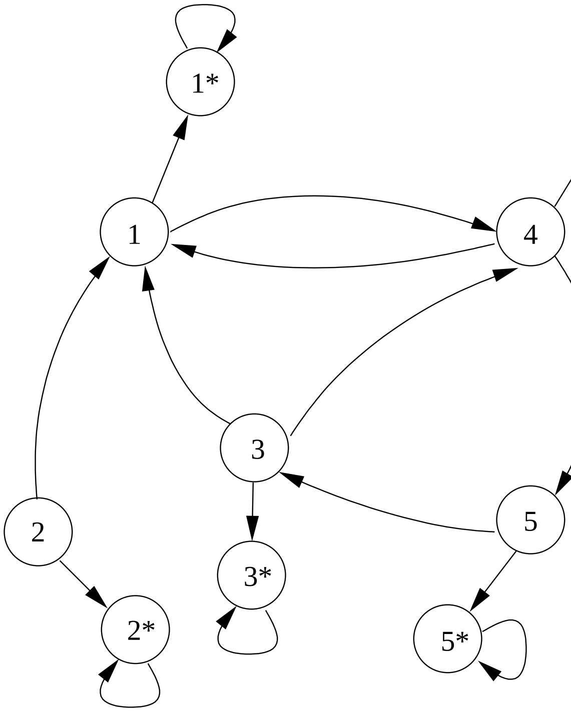

Figure 2: The extended Markov Chain including the clone states

Definition 3 Dynamic (or query dependent) mode priors: Given a query q, we create the set B of the $|B|$ top-ranked documents, where each document d_k has a score $s(d_k|q)$. In this case, the prior probability of selecting the document d_k is defined as:

$$p_k = \frac{s(d_k|q)}{2 * \sum_{d \in B} s(d|q)} \tag{19}$$

Depending on the way the prior probabilities are defined, we have two extensions of the Absorbing Model, which are described in the following sections.

3.1 Static Application of the Absorbing Model (SAM)

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:

$$s(d_j) = \sum_i p_i u_{ij^*} = \sum_i p_i u_{ij^*} = \sum_i \frac{1}{2N} p_{ij^*}^{\infty} \propto \sum_i p_{ij^*}^{\infty} \tag{20}$$

where $j^* = j + N$.

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.

To combine this link structure based score $s(d_j)$ with the score from content analysis $s(d_j|q)$, given query q, we employ the Cobb-Douglas utility function:

$$U = C^a \cdot L^b, a + b = 2 \tag{21}$$

This utility function has been applied successfully in economics to model the combination of different types of input, such as the capital and the labour. In our case, we assume that the first input is the score from content analysis $s(d_j|q)$ and the second input is the Absorbing Model score $s(d_j)$, and the parameters a, b are both set equal to 1. Consequently, the Static Absorbing Model score (SAM), where the two scores are combined, is given by:

$$RSV(d_j) = s(d_j|q) \cdot s(d_j) \tag{22}$$

3.2 Dynamic Application of the Absorbing Model (DynAMoRANK)

Another way to consider link analysis is to apply it dynamically only on the retrieved results of a first pass ranking. In this approach, the score is influenced from both the ranking and the link structure of the results. From (17) and (19), given a query q, we have for the set B of the $|B|$ top ranked documents:

$$s(d_j) = \sum_i p_i u_{ij^*} = \sum_i \frac{s(d_k|q)}{2 * \sum_{d \in B} s(d|q)} p_{ij^*}^{\infty} \tag{23}$$

where $j^* = j + N$.

The score $s(d_j)$ is a combination of both the evidence from the content analysis as well as from the link structure analysis and we set:

$$RSV(d_j) = s(d_j) \tag{24}$$

We consider the application of link analysis in this setting as an optimisation of the results returned from the first pass retrieval. Under this assumption, we ignore the outgoing links from the set A of the $|A|$ top ranked documents, before applying link analysis, where $0 \leq |A| \leq |B|$. In this way, we ensure that documents in set A only receive authority from the rest of the documents in set $B - A$. This additional restriction is justified by the fact that the evidence from the link structure is weaker than the evidence from the content of documents [9]. Therefore, we do not want to change dramatically the ranking of the first pass retrieval. We should note that the introduction of the set A does not affect in anyway the formalism of the dynamic application of the Absorbing Model, which we call DynAMoRANK.

4 Experiments and Results

In order to evaluate the effectiveness of the Absorbing Model in both its modes of application, we performed a set of experiments using standard TREC Web test collections. In the following sections, we describe the experiments and we provide an analysis of their results.

4.1 Description of experiments

For testing the Absorbing Model, we use the test collections of the Web track of TREC10 and TREC11. In particular, we employ the WT10g test collection for the topic relevance task for TREC10 and the .GOV test collection for the topic distillation and named entity tasks for TREC11. The former set of documents is a test collection that consists of 1.69 million documents [2], which were crawled in 1997. The collection comes with 50 queries for the topic relevance task, which involves finding documents about the topic of the query. The latter data set, namely the .GOV collection, is a recent crawl of the .GOV domain, that consists of 1.25 million documents. It comes with 50 topics for the topic distillation task, which is about finding documents that would serve as entries in a bookmark list for the topic of the query, or that would be linked from a portal about the the topic of the query. For the named entity finding task, there are 150 queries, for which we are looking for a single document that corresponds to what the user is searching for. In the following tables, each row is labelled with the name of the collection used for the corresponding experiments.

The indexing of the two test collections involves the removal of stop words listed in a standard stop word list [27]. Moreover, Porter's stemming algorithm [23] is applied to both collections.

For the content analysis module, acting as a baseline, we use the probabilistic framework of Amati and Van Rijsbergen [1]. We apply two term weighting functions, namely I(n_e)B2 and PL2 as introduced in [1]. For some of the experiments, the documents are augmented with the anchor text of their incoming links.

For the analysis of the link structure of the test collections, we take into account only the hyperlinks between documents in different domains. A link is used if the domain name of the URL of the source document is different than the domain name of the URL of the destination document. We apply SAM and DynAMoRANK to the topic relevance task for TREC10 and to the topic distillation task for TREC11. Moreover, DynAMoRANK is applied to the named page finding task for TREC11. For the named page finding task, the values used for $|B|$ and $|A|$ are set experimentally to 10 and 5

Notation reference	
I($\mathrm{n_e}$)B2	Content only run using the weighting function I($\mathrm{n_e}$)B2
PL2	Content only run using the weighting function PL2
+ Anchors	The anchor text is added to the main body of documents
SAM	Static Absorbing Model
PR	PageRank applied in a static mode, using the utility function (21)
DynAMoRANK	Dynamic Absorbing Model
dPR$_\mathrm{u}$	PageRank applied in a dynamic mode, using the utility function (21)
dPR$_\mathrm{b}$	PageRank applied in a dynamic mode, using the content retrieval scores to alter the transition probability matrix

Table 1: Description of the notation used in tables.

respectively.

PageRank is also compared to the Absorbing Model in both the static and the dynamic setting, although originally PageRank is designed for application in the static setting. In correspondence with SAM, the static PageRank scores are combined with the content retrieval scores using the utility function (21). For comparison with DynAMoRANK, the dynamic PageRank scores are computed only for the top $|B|$ documents of a first pass retrieval by ignoring the outgoing links from the top ranked $|A|$ documents. Then, we have two options for combining the PageRank scores with the content retrieval scores. The first is to apply the utility function (21). Alternatively, we include the content retrieval scores in the transition probability matrix, similarly to the approach taken by Haveliwala [12]. The values used for $|B|$ and $|A|$ for the dynamic link analysis are set experimentally to 50 and 20 respectively.

For enhancing the readability of results, we include Table 1 where the notation used is connected back to the description of experiments.

Content-only and anchor text experiments		
	I(n_e)B2	I(n_e)B2 + Anchors
Precision at 10 documents retrieved		
WT10g	0.3720	0.3800
.GOV	0.2408	0.2408
Mean Average Precision		
WT10g	0.2105	0.2101
.GOV	0.1990	0.1932

Table 2: Comparison of content only experiment with respect to anchor text analysis.

Content-only and anchor text experiments		
	PL2	PL2 + Anchors
Precision at 10 documents retrieved		
WT10g	0.3660	0.3560
.GOV	0.2694	0.2510
Mean Average Precision		
WT10g	0.2090	0.2065
.GOV	0.2041	0.1894

Table 3: Comparison of content only experiment with respect to anchor text analysis.

4.2 Results and Discussion

We use two different TREC test collections, which were designed having in mind different types of tasks. Indeed, the topic relevance task for the WT10g collection is closer to a classic ad-hoc task, than the topic distillation task for the .GOV collection. The latter resembles more what the actual users are expecting from a search engine as a result. For this reason, it makes more sense to use as a measure of effectiveness the precision at 10 retrieved documents, since Web users tend to browse only the top ranked documents returned by search engines.

The conducted experiments show that the selection of a weighting function affects the precision of the results. From Tables 2 and 3, it appears that the weighting function I(n_e)B2 performs better than PL2 when it is used for retrieval from the WT10g collection, while the weighting function PL2 is more suitable for use with the .GOV collection. This is an important issue when there are more than one available weighting functions and we do not know which is the most suitable. For this reason, methods for assessing automatically the effectiveness of a retrieval weighting method, such as the one pro-

Static link analysis models						
	I(n_e)B2	SAM	PR	PL2	SAM	PR
Precision at 10 documents retrieved						
WT10g	0.3720	0.3700	0.0780	0.3660	0.3700	0.0840
.GOV	0.2408	0.0082	0.0653	0.2694	0.0082	0.0796
Mean Average Precision						
WT10g	0.2105	0.2040	0.0196	0.2090	0.2027	0.0181
.GOV	0.1990	0.0187	0.0338	0.2041	0.0221	0.0931

Table 4: Comparison of link models applied in static mode.

posed by Jin et al. [16], or by Manmatha et al. [19], may be useful in selecting the most appropriate model for a first pass retrieval.

The use of anchor text exhibits different results when used in different tasks. As it appears in Tables 2 and 3, for the topic relevance task for TREC10 and the topic distillation task for TREC11, augmenting the documents with the anchor text of the incoming links seems to have a detrimental effect[1]. On the other hand, the anchor text proves to be effective when it is used for the named page finding task of the .GOV collection. In this case the average reciprocal precision increases significantly when the anchor text is incorporated in the body of the documents (Table 7). For this task the use of anchor text is even more effective than the application of pure link analysis: 0.654 w.r.t. 0.555 average reciprocal precision. We believe that the difference in retrieval effectiveness of the same source of evidence is due to the inherent differences of the nature of the topic distillation and the topic relevance compared to the named page finding task. The effectiveness of using anchor text for the first two tasks may increase if a more intelligent method is employed for adding only the anchor text that is similar to the query.

A main point of the experiments performed is the evaluation of the effectiveness of dynamic link analysis against static link analysis. It appears that DynAMoRANK and dynamic PageRank (see Tables 5 and 6) are more robust approaches than their static counterparts (see Table 4) for both the test collections we used and the weighting functions we applied. For example, for the WT10g collection, for both weighting functions, DynAMoRANK outperforms SAM with respect to the mean average precision: 0.2109 w.r.t. 0.2040 for I(n_e)B2 and 0.2072 w.r.t. 0.2027 for PL2. Moreover, it is evident from the same tables that the dynamically applied PageRank clearly outperforms the statically applied one (dPR_u and dPR_b vs. PR).

[1] The precision at 10 documents, for which we get a slight improvement from 0.3720 to 0.3800 is not the official measure of performance for the ad-hoc task for WT10g.

Content Model: I(n_e)B2				
	Baseline	DynAMoRANK	dPR$_u$	dPR$_b$
Precision at 10 documents retrieved				
WT10g	0.3720	0.3700	0.2640	0.3700
.GOV	0.2408	0.2490	0.1878	0.2490
Mean Average Precision				
WT10g	0.2105	0.2109	0.1588	0.2108
.GOV	0.1990	0.2035	0.1265	0.2045

Table 5: Comparison of link models applied in dynamic mode with I(n_e)B2 as content model.

Content Model: PL2				
	Baseline	DynAMoRANK	dPR$_u$	dPR$_b$
Precision at 10 documents retrieved				
WT10g	0.3640	0.3640	0.2620	0.3620
.GOV	0.2694	0.2776	0.1939	0.2694
Mean Average Precision				
WT10g	0.2083	0.2072	0.1545	0.2065
.GOV	0.2041	0.2047	0.1335	0.2040

Table 6: Comparison of link models applied in dynamic mode with PL2 as content model.

Looking into the results in Table 4, the performance of SAM is *a priori* intriguing. For instance, for the WT10g collection, SAM is robust for both weighting functions: 0.2040 w.r.t. 0.2105 for I(n_e)B2 and 0.2027 w.r.t. 0.2090 for PL2. However, for the .GOV collection, we notice that SAM has a clear detrimental effect. This result might suggest that the uniform distribution hypothesis underlying SAM works only for ad-hoc tasks (i.e. WT10g), while a first ranking based prior distribution seems to be necessary for the topic distillation task.

The comparison of the dynamic link analysis approaches provides us with interesting observations. From the Tables 5 and 6 it appears that DynAMoRANK outperforms both dPR$_u$ and dPR$_b$, with an exception for the case of average precision for the .GOV collection when dPR$_b$ is used with I(n_e)B2. However, the nature of the topic distillation task for .GOV suggests that we are more interested in the precision at 10 retrieved documents.

5 Conclusions

In this work we introduce a new approach for analysing the link structure of the Web and improving effectiveness of Web IR. The main problems of link analysis methods, based on random walks on the Web graph, is that the resulting scores are not always meaningful, due to the nature of the Web. For example, it is possible that some Web documents do not have any outgoing links. The approach taken by Brin and Page in PageRank [5] is to link every document to every other document, by assuming that there is a random surfer that decides at any time whether he will continue following links from the document he is visiting, or he will jump to a randomly selected document. This step is necessary to transform the corresponding Markov chain to an irreducible Markov chain, so that the scores of documents exist and they are non-zero.

We propose the Absorbing Model, in which we take a different approach in defining the corresponding Markov chain. We introduce a set of new states in the Markov chain, in a one-to-one correspondence with the states in the original Markov chain. This transformation results into a structure where the prior probabilities of selecting a document influence the probability of accessing a node. We exploit this aspect of the model in order to combine content and link analysis in a well-founded way.

The experiments we conducted show that DynAMoRANK, that is the dynamic application of the Absorbing Model, is a robust model for combining in a principled way evidence from content and link analysis. It is tested on two Web test collections, namely the WT10g and the .GOV from TREC10 and TREC11 respectively. The results show that DynAMoRANK outperforms PageRank under the same experimental settings and that it improves precision over the content-only baseline (0.2776 w.r.t. 0.2694 precision at 10 documents). This is quite important if we consider the fact that both these test collections do not favour the application of link analysis, as it has been noted in TREC10 [13] and TREC11 [8]. Thus, the results provide evidence that support the usefulness of

	$I(n_e)B2$	DynAMoRANK	$I(n_e)B2$ + Anchors
Average Reciprocal Precision			
.GOV	0.552	0.555	0.654
Pages found in top 10 documents			
.GOV	107 (71.3%)	107 (71.3%)	128 (85.3%)
Pages not found in top 50 documents			
.GOV	23 (15.3%)	22 (14.7%)	14 (9.3%)

Table 7: Named entity finding task with $I(n_e)B2$.

link analysis as a method for increasing precision among the top ranked documents, a still open issue for the field of Web information retrieval.

Finally, the Absorbing Model is a flexible and theoretical approach, where the combination of content and link analysis is achie-ved by using the priors, without increasing the computational cost. For ad-hoc retrieval, the use of the uniformly distributed priors is robust, while the topic distillation task is topic-driven, which might suggest that the priors affect the retrieval effectiveness. Further investigation is required to relate the nature of the task to the importance of the priors. The results show that the Absorbing Model is a promising method that could improve precision over the content-only baseline experiments.

References

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[2] P. Bailey, N. Craswell, and D. Hawking. Engineering a Multi-Purpose Test Collection for Web Retrieval Experiments. Accepted by the Information Processing and Management journal.

[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] A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *Proceedings of the tenth international conference on World Wide Web*, pages 415–429. ACM Press, 2001.

[5] 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.

[6] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proceedings of the tenth international conference on World Wide Web*, pages 211–220. ACM Press, 2001.

[7] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Proceedings of the 24th annual international ACM*

SIGIR conference on Research and development in information retrieval, pages 208–216. ACM Press, 2001.

[8] N. Craswell and D. Hawking. Draft overview of the trec 2002 web track, 2002.

[9] W.B. Croft. Combining approaches to information retrieval. In W.B. Croft, editor, *Advances in Information Retrieval from the Center for Intelligent Information Retrieval*. Kluwer Academic, 2000.

[10] W. Feller. *An Introduction to Probability Theory and its Applications, volume 1, 2nd edition*. John Wiley and Sons, 1957.

[11] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.

[12] T.H. Haveliwala. Topic-Sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.

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

[14] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM Press, 1999.

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

[16] Rong Jin, Christos Falusos, and Alex G. Hauptmann. Meta-scoring: automatically evaluating term weighting schemes in ir without precision-recall. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–89. ACM Press, 2001.

[17] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[18] 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.

[19] 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.

[20] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[21] A.Y. Ng, A.X. Zheng, and M.I. Jordan. Stable algorithms for link analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 258–266. ACM Press, 2001.

[22] F. Osareh. Bibliometrics, citation analysis, and co-citation analysis: A review of literature i. *Libri*, 46:149–158, 1996.

[23] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[24] 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.

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

[26] Ilmrio Silva, Berthier Ribeiro-Neto, Pvel Calado, Edleno Moura, and Nvio 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.

[27] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworth-Heinemann, 1979.