# Cluster Enhanced Pageranks

Ankur Gupta
ankur@cs.utexas.edu

Sreangsu Acharyya
srean@lans.ece.utexas.edu

The University of Texas at Austin

December 15, 2003

**Abstract**

Abundance of pages on the web containing common query terms, is one of the biggest challenge for a web search engine, and has led to the development of link analysis algorithms. Pagerank, developed and used by Google is one of the most well known such algorithms, and operates by assigning values to the vertices of the webgraph that can be interpreted as the stationary distribution of a Markov chain defined on them. The Markovian surfing behavior consists of two components chosen stochastically, (i) where the surfer samples an outlink uniformly at random and (ii) another where the surfer jumps uniformly to a random page of the web. The second behavior is independent of any property of the graph and had been introduced in order to safeguard against the possibility of pagerank values of pages(vertices) taking a value of 0 and to ensure a unique stationary distribution, by assuring that the Markov chain is irreducible. Its the second term which is the target of our interest here, we propose an alternative behavior that ensures the same condition but is based on the property of the graph rather than being completely random as in the original pagerank algorithm. We look at a coarser representation of the web, by considering transitions between subsets of vertices identified by a suitable clustering algorithm. The random mode now consists of picking a subset according to the coarsened Markov chain and then choosing a vertex from that subset based on a well specified distribution. An outcome of such a formulation is that the pagerank of a page is dictated not only by the quality of inlinks but also by its membership to a cluster, thereby allowing a new, "*inlink scarce*" page to have a high pagerank if it belongs to a "good" cluster. For the current implementation we have chosen the clustering based on the outlinks of a page, but other possibilities such as content need to be examined. We report experiments on the University of Texas webgraph and the Cora citation graph to provide anecdotal evidence.

## 1 Introduction

Selecting and ordering query results, from over 3 billion hyper-linked pages that now constitutes the web graph $G(V, E)$, is a difficult web mining problem of extreme importance, made particularly difficult by the abundance of web pages containing common query terms. Link analysis is one of the key techniques that has made effective solutions possible. These techniques have been motivated by the view that the web can be looked upon as a noisy, distributed recommendation system where each page recommends other pages through its outlinks. This leads to a possible representation of the local recommendation as a $|V|$ dimensional vector where the value at the $i^{th}$ location (possibly binary) denotes the strength of recommendation. Thus the important task at hand is to collate these local recommendations of varying credibility into a global one. A simple strategy is to use a weighted linear combination of the "recommendation vectors" of all pages. This still keeps an important question unanswered: what should the weights of combination be. An elegant choice is one where the weights are the same as that of the final global recommendation weights. Implementing this

idea, using recommendation vectors that have been $L_1$ normalized leads to the Pagerank algorithm, which is described in detail later in the section to follow. The final global recommendation vector in this setting is called the Pagerank vector, and forms one of the key measures used by the popular search engine Google to rank search results.

# 2 Pagerank

The weighted linear combination of local recommendations as computed by the Pagerank algorithm is equivalent to ranking web pages according to the stationary distribution of a random surfer traversing the directed web graph $G(V, E)$. The surfer is assumed to be Markovian and traverse the web graph mostly by following outlinks, with occasional jumps to a random page on the web. The surfer either picks one outlink from the current page uniformly at random or resets to a random page on the web with probabilities $\alpha$ and $1 - \alpha$ respectively. The escape probability of $(1 - \alpha)$ was introduced as a mechanism to ensure non-zero Pageranks and irreducibility of the resulting Markov chain [1] which is required for the Markov chain to have an unique stationary distribution.

Let $A$ be the adjacency matrix of the graph, $D_{out}$ the diagonal matrix formed by the out-degrees, $S$ a row stochastic source matrix usually taken to be $S = \frac{1}{N}(e \times e^T)$, where $e$ is a column vector of ones. The transition matrix $T$ of the surfer can be expressed in terms of the out-degree normalized adjacency matrix $D_{out}^{-1}A$, and the random jump probability $\alpha$ as :

$$T = \alpha \times D_{out}^{-1} \times A + (1 - \alpha) \times S \tag{1}$$

The Pagerank iteration converges to the primary eigenvector $\pi$ of $T^T$ or, equivalently, the steady-state probabilities of the Markov chain defined by it. The convergence rate is proportional to $\lambda_1/\lambda_2$, the ratio of the first and second eigenvalues of $T$ [1].

## 2.1 Random Jumps

One of the main objective of adding rank source matrix $S$ in the original Pagerank algorithm is to avoid the rank sink problem, more specifically handling the presence of absorbing states by making the Markov chain irreducible. A set of states $\mathcal{V}_S \subset V$ is called *absorbing* if $P(\overline{\mathcal{V}_S} \to \mathcal{V}_S) > 0$ but $P(\mathcal{V}_S \to \overline{\mathcal{V}_S}) = 0$. Note that in addition to absorbing states one might have sets of isolated states $\mathcal{V}_I$ such that $P(\overline{\mathcal{V}_I} \to \mathcal{V}_I) = P(\mathcal{V}_I \to \overline{\mathcal{V}_I}) = 0$.

Among these its the presence of absorbing states which prevent meaningful values for Pageranks since the pageranks of all non-absorbing pages go to zero, whereas the presence of isolated states can be handled suitably by proper initialization and considering the decoupled chains separately. The most common form of the Matrix $S$ is $ee^T$ though $S = se^T$ has been utilized for personalizing search results in [2].

We show that it is possible to remove absorbing states using a technique other than uniform random jumps to arbitrary pages. Consider the case where a random surfer visits a page $i$, jumps with probability $\alpha$ to one its outlinks uniformly at random and with probability $(1-\alpha)$ *it jumps to one if its inlinks uniformly at random.* It can be proved that the resulting walk contains no absorbing states and under the assumption that no isolated states exist, the resulting Markov Chain is irreducible. Let us refer to the stationary distribution of this walk under uniform initialization as the "forward-backward ranks". Some properties of these ranks along with methods for choosing the value of $\alpha$ were discussed in [3].

**Lemma 1** *A Forward Backward Markov chain cannot have absorbing states.*

However, a significant drawback of this approach is that it is vulnerable to spam. A web page can increase its Pagerank just by adding outlinks. To reduce this, we first cluster the vertices of the graph by using an

appropriate clustering algorithm and then use it to compute the "forward-backward" ranks for the blocks.[1] The columns of the matrix $S$ are taken to be the weighted concatenation of the local forward-backward ranks, where the weights are given by the block ranks. Note that unless all nodes in a cluster collude together, it is hoped that this approach is resistant to spam. Coarser representation of the web has been used by [4] in order to speedup the pagerank calculation, however there the authors did not investigate clustering algorithms, as reverse lexicographical sorting of the urls led to a satisfactory block structuring the adjacency matrix.

# 3 Algorithm: Clustered Pagerank

We modify the graph independent part of the pagerank surfer's behavior so that it does not visit pages arbitrarily. This is constructed as a two stage process as follows: suppose we partition/cluster the set of vertices of the webgraph into $K$ clusters according to some similarity criteria (we have used the information-theoretic co-clustering algorithm as described in [5]), now instead of selecting an arbitrary page from the web uniformly at random, the surfer first selects a cluster to transit to depending on its current page and continues a random walk on the cluster labels till convergence, after which it chooses a particular member from its current cluster.

From the property of Markov chains we know that as long as the transitions on the cluster labels are irreducible the initial distribution over the cluster label is of no consequence as far as the converged distribution is concerned. Thus inorder to completely define the behavior of the surfer we need to specify (i) the distribution over clusters, which guides the surfer's choice of the cluster label to transit to (essentially the stationary distribution of the random walk on the cluster labels) and (ii) the distribution over the members of the cluster which dictates the choice of a particular page having chosen a cluster label. Both these quantities are defined in terms of the property of the graph.

First let us consider constructing the transition matrix of the Markov chain on the $K$ cluster labels, whose stationary distribution we need to compute. The $(I, J)^{th}$ entry of the transition matrix $T_f$ indicates the probability of visiting any vertex in the $J^{th}$ cluster conditioned on the fact that the current vertex is in $I$. This is a well defined quantity given any initial distribution $R$ over set of vertices and equals

$$\frac{\sum_{i\in I, j\in J} R[i](D_{out}^{-1}A[i][j])}{\sum_{i\in I} R[i]} = \sum_{i\in I, j\in J} (D_{out}^{-1}A[i][j]) \left(\frac{R[i]}{\sum_{i\in I} R[i]}\right).$$

Note that we can use any locally normalized distribution instead of $\left(\frac{R[i]}{\sum_{i\in I} R[i]}\right)$ and it would corresponds to some initialization vector, for our implementation we use the original pagerank to be the initialization vector.

A reasonable choice for second distribution that we need to specify, the distribution over the members of a cluster which dictates the choice of a particular page having chosen a cluster label, is the local pageranks computed per cluster. However for implementation purpose we used the locally normalized global pagerank values computed from the original algorithm.

**Variables**
$G = (V, E)$ : webgraph;
$N$ : no. of webpages;
$K$ : no. of clusters;

[1] Henceforth, we will use cluster and block interchangeably.

$i, j$ : webpages;
$I, J$ : clusters;
$A$ : Adjacency matrix of the web graph;
$D_{out}$ : Diagonal out degree matrix if $G$ s.t. $D_{out}[i,i] = Outdegree(i)$.
$D_{in}$ : Diagonal in degree matrix if $G$ s.t. $D_{in}[i,i] = indegree(i)$.
$T_b$ : $K \times K$ Reverse Block Transition Matrix;
$T_f$ : $K \times K$ Forward Block Transition Matrix;
$LPR_I[j]$ : Local Rank of page $j$ in block/cluster $I$ ;
$PR[j]$ : Final global Pagerank of page $j$ ;
$\beta$ : damping factor for computing Block Rank Vector;
$B$ : Block Rank Vector;
$S$ : $N \times N$ source matrix;

**Algorithm**
[1] Cluster the graph based on some clustering algorithm.
Compute local pageranks for each cluster.
Compute the Block Transition Matrices as follows:
$T^f[I][J] = \sum_{i \epsilon I,\ j \epsilon J} (D_{out}^{-1} A)[i][j] \times LPR_I[i]$
$T^r[I][J] = \sum_{i \epsilon I,\ j \epsilon J} (D_{in}^{-1} A^T)[i][j] \times LPR_I[i]$
Compute Block Rank Vector $B$ as follows:
$B = Dominant\ Eigen\ Vector\ of\ (\beta \times T^{fT} + (1-\beta) \times T^{rT})$
Compute the source matrix $S$ as follows:
$S[i][j] = B[J] \times LPR_J[j]\ where\ j \epsilon J$
Compute the final global Pagerank vector $PR$ as follows:
$PR = Dominant\ Eigen\ Vector\ of\ (\alpha \times A^T + (1-\alpha) \times S^T)$

**Lemma 2** *The block transition matrices are row-stochastic.*

**Proof.** We show that the Forward Block Transition Matrix is row-stochastic. The other one can be shown in a similar fashion.
$T^f[I] = \sum_J \sum_{i \epsilon I,\ j \epsilon J} (1/D_{out}[i][i]) \times A[i][j] \times LPR_I[i]$
$= \sum_{i \epsilon I} LPR_I[i] \times \sum_j (1/D_{out}[i][i]) \times A[i][j]$
$= \sum_{i \epsilon I} LPR_I[i]$
$= 1$ {Local ranks are normalized to 1}

□

**Lemma 3** *The rank source matrix $S$ is row-stochastic.*

**Proof.** Proof is similar to that in Lemma 2. □

**Theorem 1** *Assuming there are no isolated blocks in the clustered web graph, the above algorithm ensures that the matrix $\alpha \times T^T + (1-\alpha) \times S^T$ is irreducible.*

We explain the differences in ranking that would be produced by the original Pagerank algorithm and the modified version with a figure (see Figure 1). Consider that a finite rank (probability) mass is distributed among the vertices proportional to the original Pagerank algorithm. In the original Pagerank algorithm, each vertex distributes a preassigned constant ($\alpha = 0.85$) fraction of its share of the mass equally among the pages to which it links and the rest of the mass equally among all the pages on the web. On the receiving side, each page receives a fraction of their parents ranks and a *constant fraction* $(1-\alpha)$ of the ranks from all

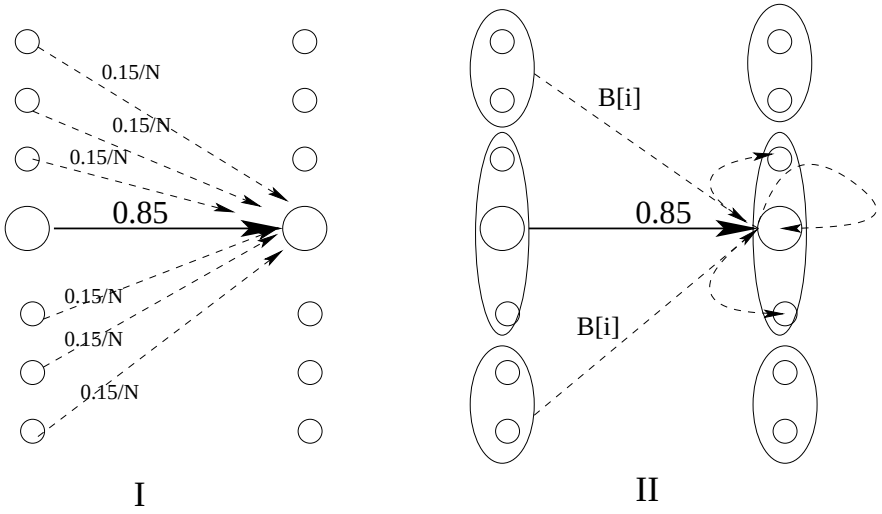


Figure 1: Pagerank vs Cluster Enchanced Pagerank: Both the figures I and II show a bipartite view of the graph, where vertices drawn to the left link to vertices drawn to the right. Fig 1.I shows a vertex receiving 0.85 rank mass from its parent and $0.15/N$ rank mass from all vertices as in the original Pagerank algorithm. Fig 1.II shows a vertex belonging to cluster $i$ receive rank mass from its parents and cluster label dependent rank mass from other pages as in the modified algorithm

pages. In the modified algorithm a donor page distributes the same fraction ($\alpha = 0.85$) of its rank among its children as before, however the remaining mass is distributed preferentially among the clusters proportional to the block rank $B[.]$ and eventually to its members proportional to their local rank $LPR$. Considering the recipient point of view, each page receives a fraction of their parents mass as before but from all the pages on the web it receives a *non constant* fraction of their mass which depends on the block rank $B[.]$ cluster label of the recipient page and the position of the recipient page within the cluster. Thus for a recipient page $i$ with cluster label $I$ if

$$B[I] * LPR[i] > \frac{(1-\alpha)}{N}$$

the net rank mass it receives is higher than the standard pagerank algorithm as a result of which its modified pagerank would also be higher. The proof of the above requires a lemma about the monotonicity property of the stationary distribution of a Markov chain proved recently in [6].

**Lemma 4** *[6] Let $P$ be a finite-state ergodic Markov chain and let $a$, $b$ be arbitrary states of $P$ and $c \neq a$. Let $\bar{P}$ be a Markov chain obtained by increasing the transition probability to $a$ such that $\bar{p}_{ba} = p_{ba} + \delta$ and reducing the transition to some $c$ such that $\sum_{c \neq a} \bar{p}_{bc} = \sum_{c \neq a} p_{bc} - \delta$. Then if $\bar{P}$ has a stationary distribution $\bar{\pi}$ then $\bar{\pi}_a > \pi_a$.*

**Theorem 2** *For the same value of $\alpha$ if $I(i)$ is the cluster label of the page $i$ and $B[I(i)] * LPR[i] > \frac{(1-\alpha)}{N}$ holds then the rank of the page $i$ under the modified algorithm would be higher.*

**Proof.** We showed that one step of the modified algorithm assigns more rank mass to a page if the above condition is met. For the full proof one needs to use the monotonicity property of stationary distribution of a Markov chain mentioned above. □

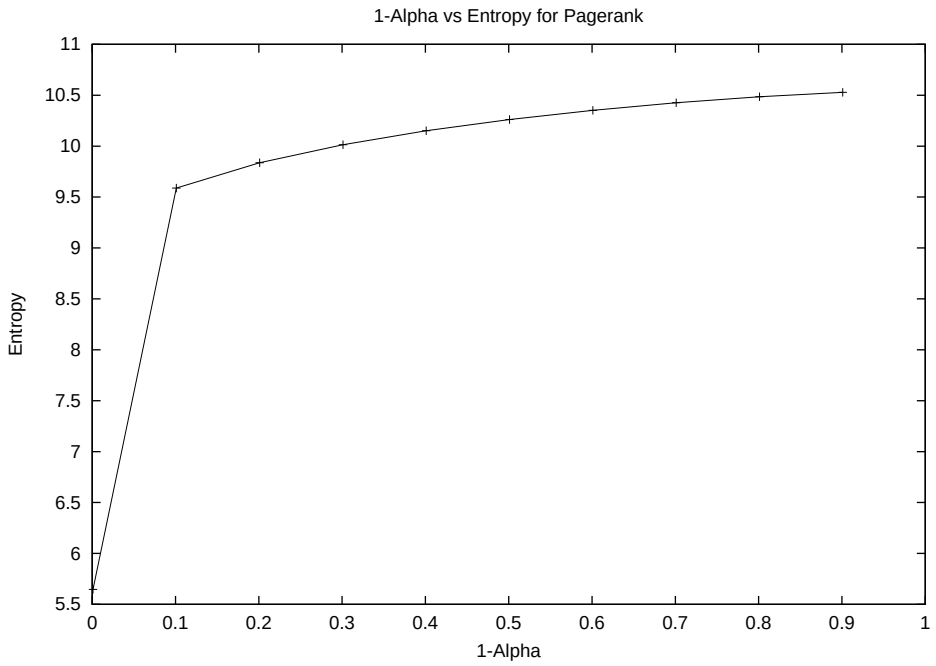


Figure 2: $1-\alpha$ *vs Entropy*

# 4 Choosing $\alpha$

Choosing the value of $\alpha$ is a big challenge in the Pagerank algorithm (and all its derivatives). In absence of any information, the Minimum Assumption Principle says that the entropy of the Pageranks obtained must be maximum. This suggest using as small a value of $\alpha$ as possible. But we also want that the contribution from link structure of the web graph be as much as possible. This requires choosing $\alpha$ as large as possible. We, therefore, looked at how the entropy varies with $\alpha$ and tried to find a value of $\alpha$ that is a sweet-spot for these conflicting requirements.

**Lemma 5** *$H[\alpha A^T r + (1-\alpha)S^T r]$ monotonically decreases as $\alpha$ increases, where $H$ is the entropy.*

**Proof.** Let $H{=}H[\alpha A^T r + (1-\alpha)S^T r]$
$H = -[\sum_{i=1..n}(\alpha c_i^T r + (1-\alpha)\frac{1}{n})\ log(\alpha c_i^T r + (1-\alpha)\frac{1}{n})]$ $\{A = [c_1\ c_2\ ...\ c_n]\}$
$\frac{dH}{d\alpha} = -[\sum_{i=1..n}(c_i^T r - \frac{1}{n})\ log(\alpha c_i^T r + (1-\alpha)\frac{1}{n})] + \sum_{i=1..n}(c_i^T r - \frac{1}{n})$
$= -[\sum_{i=1..n}(c_i^T r - \frac{1}{n})\ log(\alpha c_i^T r + (1-\alpha)\frac{1}{n})]]$ {Second term is zero as $A$ is row-stochastic.}
$\frac{d^2H}{d\alpha^2} = -[\sum_i \frac{((c_i^T r - \frac{1}{n})^2}{(\frac{1}{n} + \alpha(c_i^T r - \frac{1}{n}))}] < 0$

Therefore, $\frac{dH}{d\alpha}$ monotonically decreases with maximum being at $\alpha = 0$. At $\alpha = 0$, $\frac{dH}{d\alpha} = 0$. Therefore, $H$ decreases with $\alpha$. □

**Conjecture 1** *Entropy of Pagerank monotonically increases as $\alpha$ decreases.*

**Proof.** From Lemma 5 it seems that the conjecture holds. Empirical evidence also vindicates this. We ran the original Pagerank algorithm for 1500 docs obtained from a webcrawl of *utexas.edu* domain. The

| No. of nodes | No. of clusters | OSim | KSim |
|---|---|---|---|
| 1000 | 10 | 0.9 | 0.747368 |
| 1000 | 20 | 0.9 | 0.847368 |
| 1000 | 30 | 0.9 | 0.757895 |
| 1000 | 40 | 0.9 | 0.847368 |
| 1000 | 50 | 0.9 | 0.857895 |
| 1500 | 10 | 0.85 | 0.642105 |
| 1500 | 20 | 0.85 | 0.642105 |
| 1500 | 30 | 0.85 | 0.642105 |
| 1500 | 40 | 0.85 | 0.642105 |
| 1500 | 50 | 0.85 | 0.642105 |
| 2000 | 10 | 0.85 | 0.6 |
| 2000 | 20 | 0.85 | 0.594737 |
| 2000 | 30 | 0.85 | 0.594737 |
| 2000 | 40 | 0.85 | 0.594737 |
| 2000 | 50 | 0.85 | 0.568421 |
| 2500 | 10 | 0.85 | 0.578947 |
| 2500 | 20 | 0.85 | 0.578947 |
| 2500 | 30 | 0.85 | 0.573684 |
| 2500 | 40 | 0.85 | 0.584211 |
| 2500 | 50 | 0.85 | 0.584211 |

Table 1: OSim and KSim values for the top 20 pages.

variation of *Entropy* as a function of $1-\alpha$ is given in Figure 2. As can be seen, the entropy monotonically increases as $\alpha$ decreases. □

From Figure 2, we see that[2] $\alpha = 0.85$ (i.e. $1-\alpha = 0.15$) is actually a good choice as far as maximizing entropy is concerned. Hence in all of our experiments we use this value. To simplify things, we take $\beta$ also as 0.85.

# 5 Experiments

Experiments were performed on 2 datatsets using the information theoretic co-clustering to cluster the pages. The information theoretic co-clustering is technique for quantizing a joint distribution, which was calculated from the transition probabilities and its corresponding stationary distribution. In the first set, we worked on a webcrawl of *utexas.edu* that consisted of more than 30,000 pages. We studied how the Pageranks change on subgraphs consisting of 1000-2500 pages and number of clusters varying between 10-50 for both the original pagerank scheme and our modified scheme. In search for a good metric for evaluation, we settled for OSim and KSim measures as defined and used in [2]. Table 1 shows the variation in OSim and KSim values between the original Pagerank and our algorithm for different subgraph sizes and clusters. The similarity measures were calculated for the top 20 pages. The results show that the two algorithms produce quite similar results. This tends to indicate that probably its not worth the effort adding the extra complexity to the Pagerank algorithm. We also studied how the entropy of the Pageranks produced by the original

[2] Note that for 1500 docs, the maximum entropy possible is $\sim 10.55$.

| **Original** | **New** |
|---|---|
| www.utexas.edu/ | www.utexas.edu/ |
| www.bus.utexas.edu/ | www.utexas.edu/teamweb/ |
| www.utexas.edu/teamweb/ | www.bus.utexas.edu/ |
| www.utexas.edu/homecomments.html | www.utexas.edu/homecomments.html |
| www.utexas.edu/directory/offices/ | www.utexas.edu/directory/offices/ |
| x500.utexas.edu/ | x500.utexas.edu/ |
| www.utexas.edu/search/ | www.utexas.edu/index.html |
| www.psy.utexas.edu/ | www.utexas.edu/search/ |
| www.csr.utexas.edu/ | www.utexas.edu/do/ |
| www.tsgc.utexas.edu/graphics/buttons/www.map | www.tsgc.utexas.edu/graphics/buttons/www.map |
| www.utexas.edu/help/ | www.csr.utexas.edu/ |
| www.utexas.edu/index.html | www.utexas.edu/help/ |
| www.utexas.edu/do/ | www.psy.utexas.edu/ |
| dev.lib.utexas.edu/services/reference/email.html | dev.lib.utexas.edu/services/reference/email.html |
| www.biosci.utexas.edu/mgm/ | www.tsgc.utexas.edu/ |
| www.tsgc.utexas.edu/ | www.cps.utexas.edu/ |
| www.cps.utexas.edu/ | www.utexas.edu/utpostal.html |
| www.tsgc.utexas.edu/info/teamweb.html | www.tsgc.utexas.edu/info/teamweb.html |
| www.cm.utexas.edu/services/computer/webadmin.html | ccwf.cc.utexas.edu/ informs/links.html |
| www.utexas.edu/ogs/ | www.cps.utexas.edu/home.html |

Table 2: Pagerank vs Cluster Enhanced Pagerank Urls

Pagerank compares with ours. Results show that the original Pagerank achieves 93% compared to 88% (of the maximum entropy possible) of our algorithm. This result is expected as the personalization vector used by our algorithm has less entropy (since it is more biased towards important pages) than the uniform vector chosen in the original scheme.

The second set of experiments that we carried out were on a subset of the cora dataset consisting of the citation graph of 11754 research papers in the "Artificial Intelligence" category. Cora was a popular search engine for computer science research papers, and has previously been used by Ng etal in [7] to examine the properties of the rankings produced by HITS [8] and Pagerank. We ran the Clustered pagerank algorithm on the subset with different cluster sizes, here we present ( see tables 3, 4 )the ranks occupied by the top 20 pages in the runs of the algorithm started with different clusters sizes. Unlike the utexas webgraph, here the two algorithms differ significantly in their rankings, where in the modified algorithm the top pages are dominated by the pages with high block rank.

# 6 Conclusions

In this report we presented a modification of the Pagerank algorithm which removes the graph independent behavior of the equivalent random surfer and substitutes a component which still ensures irreducibility of the Markov chain, a requirement for it to have an unique stationary distribution. The modification is based on a coarser representation of the web invoked by partitioning the pages into subsets or clusters. The clustering induces a Markov chain in as many states as clusters the stationary distribution of which assigns a block rank to each of these clusters. The *absorption avoiding* component consists of selecting a cluster proportional to the block rank values and then selecting a page from that cluster based on the local ranks. The rank of a page now depends not only on its inlinks but also by the block rank of the cluster to which it belongs. This

10 Clusters

| Rank | page id | Pagerank | Block Rank |
|---|---|---|---|
| 1 | (Valente et. al., 96) An intelligent system for air campaign plan evaluation based on EXPECT | 4.868e-01 (2) | (0) 8.937e-01 |
| 2 | (Valente et. al., 96) A Representation and Library for Objectives in Air Campaign Plans | 4.868e-01 (1) | (0) 8.937e-01 |
| 3 | (Barto et al., 1983) | 4.149e-05 (8) | (1) 2.683e-02 |
| 4 | (Barto93) realtime-dp | 3.774e-05 (12) | (1) 2.683e-02 |
| 5 | (Sutton88) Learning to predict by the methods of temporal differences | 3.537e-05 (6) | (5) 1.402e-03 |
| 6 | (Koh)Self-Organization and Associative Memory | 2.970e-05 (10) | (1) 2.683e-02 |
| 7 | (John94) | 2.798e-05 (17) | (1) 2.683e-02 |
| 8 | 777(McAllister and Rosenblitt 1991)Systematic nonlinear planning | 2.512e-05 (13) | (0) 8.937e-01 |
| 9 | 117(Szepesvari,Littman 97) Generalized Markov Decision Processes | 2.200e-05 (25) | (1) 2.683e-02 |
| 10 | 1388 (Sim94) A robust layered control system for a mobile robot | 2.128e-05 (5) | (7) 1.198e-03 |
| 11 | 28350 (Singh 94) On the convergence of stochastic iterative dynamic programming algorithms | 1.873e-05 (55) | (1) 2.683e-02 |
| 12 | 7632 Connectionist Speech Recognition: A Hybrid Approach | 1.786e-05 (29) | (2) 1.938e-02 |
| 13 | 763 (Barrett and Weld, 1992) Partial Order Planning: Evaluating Possible Efficiency Gains | 1.769e-05 (21) | (2) 1.938e-02 |
| 14 | 6163 (Boyan and Moore, 1995) Generalization in reinforcement learning: Safely approximating the value function | 1.493e-05 (64) | (1) 2.683e-02 |
| 15 | 6209 (Singh, 1992) TRansfer of Learning by Composition | 1.468e-05 (68) | (1) 2.683e-02 |
| 16 | 7423(Dean93) | 1.447e-05 (41) | (1) 2.683e-02 |
| 17 | 6214 (Sutton, 1990) | 1.4003e-05 (20) | (5) 1.402e-03 |
| 18 | 28504 (Sut95) TD models: modeling the world at a mixture of time scales | 1.285e-05 (111) | (1) 2.683e-02 |
| 19 | 35 (Goldberg) Genetic Algorithms in Search, Optimization and Machine Learning | 1.276e-05 (3) | (3) 1.560e-03 |
| 20 | 10798 ftp://ftp.cs.utexas.edupub/mooney/papers/ioueither-bkchapter-93.ps.Z | 1.273e-05 (60) | (0) 8.937e-01 |

Table 3: The Pagerank values of the top 20 pages ranked according to the modified algorithm run on the cora data set with 10 clusters is shown. The third column shows the pagerank according to the modified algorithm and its original rank in parantheses. The third column shows the block rank of the cluster to which the page belonged, the value in the parantheses shows the rankings of the clusters according to the Block rank values. Note that pages belonging to highly ranked clusters improved their rankings as compared to the original ranks given in column 3

allows fresh pages to be highly ranked provided they are members of "good clusters".

We were also able to characterize exactly when would a page be ranked differently from the conventional pagerank algorithm, using a recently proven result on the monotonic behavior of the stationary distribution of a Markov chain.

On the webgraph data set the new and the old rankings were mostly similar whereas on the citation

20 Clusters

| Rank | page id | Pagerank | Block Rank |
|---|---|---|---|
| 1 | 571354 (Valente et al., 1996) an intelligent system for air campaign plan evaluation based on EXPECT | (2) 4.772e-01 | (0) 4.055e-01 |
| 2 | 538295 (Valente et. al., 96) A Representation and Library for Objectives in Air Campaign Plans | (1) 4.767e-01 | (1) 4.001e-01 |
| 3 | 1719 (Agrawal et al. 93) Mining Association Rules between Sets of Items in Large Databases | (18) 1.259e-04 | (2) 6.350e-02 |
| 4 | 910 (poggio89) A Theory of Networks for Approximation and Learning | (22) 1.123e-04 | (2) 6.350e-02 |
| 5 | 9812 (Agrawal et al) An Interval Classifier for Database Mining Applications | (32) 1.001e-04 | (2) 6.350e-02 |
| 6 | 3925 (Knoblock, 1991) http://www.isi.edu/sims/papers/91-thesis.ps | (61) 7.071e-05 | (2) 6.350e-02 |
| 7 | 4423 (Lin94]) State Constraints Revisited | (86) 6.946e-05 | (2) 6.350e-02 |
| 8 | 184 (Cheeseman89) Where the Really Hard Problems Are | (94) 5.343e-05 | (2) 6.350e-02 |
| 9 | 28957 (Buntine, 1994)jair/pub/volume2/buntine94a.ps.Z | (117) 4.997e-05 | (2) 6.350e-02 |
| 10 | 4330 (Jordan94) Hierarchical Mixtures of Experts and the EM Algorithm | (16) 4.890e-05 | (3) 1.761e-02 |
| 11 | 9672 (Knoblock etal 91) Integration of Abstraction and Explanation-Based Learning in PRODIGY | (170) 4.703e-05 | (2) 6.350e-02 |
| 12 | 58527 (Pinto94) Temporal Reasoning in the Situation Calculus | (282) 4.537e-05 | (2) 6.350e-02 |
| 13 | 23796 (StAmant94) A planning representation for automated exploratory data analysis | (387) 4.521e-05 | (2) 6.350e-02 |
| 14 | 4348 (Knoblock90) Learning Abstraction Hierarchies for Problem Solving | (257) 4.433e-05 | (2) 6.350e-02 |
| 15 | 1272 (Quinlan86) Induction of Decision Trees | (11) 4.219e-05 | (8) 1.451e-03 |
| 16 | 1388 (Sim94) A robust layered control system for a mobile robot | (5) 3.980e-05 | (9) 1.443e-03 |
| 17 | 12851 (Crawford, et. al. 90) QPC : A Compiler from Physical Models into Qualitative Differential Equations | (237) 3.964e-05 | (2) 6.350e-02 |
| 18 | 3236 (KJL + 94) A machine learning library in C++ | (245) 3.920e-05 | (2) 6.350e-02 |
| 19 | 3584 (BvHH + 89) A rational reconstruction and extension of recursion analysis | (541) 3.798e-05 | (2) 6.350e-02 |
| 20 | 33231 (Smyth et al., 1997) ftp://psyche.mit.edu/pub/jordan/hmm.ps.Z | (278) 3.677e-05 | (2) 6.350e-02 |

Table 4: The Pagerank values of the top 20 pages ranked according to the modified algorithm run on the cora data set with 20 clusters is shown. The third column shows the pagerank according to the modified algorithm and its original rank in parentheses. The third column shows the block rank of the cluster to which the page belonged, the value in the parentheses shows the rankings of the clusters according to the Block rank values. Note that the top ranked pages mostly consist of pages with high block rank.

graph significant differences were observed. It should also be observed that the quality of the rankings is subject to the quality of the clustering obtained. For the purpose of this study the clustering was based on link-structure rather than based on contents of webpages. We believe that a content-based clustering will give superior results.

Pagerank

| Rank | page id | Pagerank | Cluster Ranks Rank |
|---|---|---|---|
| 1 | 538295 (Valente et. al., 96) A Representation and Library for Objectives in Air Campaign Plans | 3.366e-1 | 2(0), 2(1), 2(1), 2(1), 2(1) |
| 2 | 571354 (Valente et al., 1996) an intelligent system for air campaign plan evaluation based on EXPECT | 3.366e-1 | 1(0), 1(0), 1(0), 1(0), 1(0) |
| 3 | 35 (Goldberg) Genetic Algorithms in Search, Optimization and Machine Learning | 3.921e-4 | 19(3), 66(9), 140(7), 26(19), 3(3) |
| 4 | 2 (Gar) Computers and Intractibility: A Guide to the Theory of NP-Completeness | 2.810e-4 | 32(3), 3(2), 26(11), 120(7), 27(16) |
| 5 | 1388 (Sim94) A robust layered control system for a mobile robot | 2.7661-4 | 10(7), 5(1), 16(9), 403(18), 46(12) |
| 6 | 6213 (Sutton88) Learning to predict by the methods of temporal differences | 2.614e-4 | 5(5), 803(11), 516(8), 7(5), 112(9) |
| 7 | 3229 (J. Hertz & Palmer 91) Introduction to the Theory of Neural Computation | 1.701e-4 | 102(5), 424(9), 511(13), 16(6), 153(8) |
| 8 | 4584 (Barto et al., 1983) | 1.456e-4 | 3(1), 281(6), 515(8), 409(10), 129(7) |
| 9 | 2027 (Boyer and Moore 1988) A Computational Logic Handbook | 1.435e-4 | 508(6), 818(8), 90(4), 5(0), 19(0) |
| 10 | 1365 (Koh) Self-Organization and Associative Memory | 1.351e-4 | 6(1), 44(13), 528(6), 3(2), 123(7) |
| 11 | 1272 (Quinlan86) Induction of Decision Trees | 1.313e-4 | 70(5), 29(13), 15(8), 1062(17), 76(7) |
| 12 | 114 (Barto93) realtime-dp | 1.157e-4 | 4(8), 1909(11), 693(17), 400(5), 928(9) |
| 13 | 777 (McAllister and Rosenblitt'91) Systematic nonlinear planning | 1.111e-4 | 8(0), 27(6), 547(7), 1382(13), 1330(18) |
| 14 | 9890 (Cover, Thomas'91) Elements of Information Theory | 1.102e-4 | 45(3), 905(10), 692(8), 9(4), 709(11) |
| 15 | 887 (Samuels, 1959) Some studies in machine learning using the game of Checkers | 1.048e-4 | 55(8), 1217(9), 597(7), 4(2), 906(12) |
| 16 | 4330 (Jordan94) Hierarchical Mixtures of Experts and the EM Algorithm | 1.046e-4 | 23(9), 18(10), 10(3), 1355(17), 61(13) |
| 17 | 3231 (Smyth etal, 1997) hmmpin.ps.Z | 1.0114-4 | 7(1), 957(6), 561(10), 913(6), 903(10) |
| 18 | 1719 (Agrawal et al. 93) Mining Association Rules between Sets of Items in Large Databases | 9.996e-5 | 2793(5), 7(8), 3(2), 1698(22), 1436(18) |
| 19 | 783 (Penberthy92) Complete, Partial-Order Planner for ADL | 9.790e-5 | 415 (7), 12(4), 709(6), 1673(21), 1391(25) |
| 20 | 6214 (Sutton, 1990) ftp.cs.umass.edu/pub/anw/pub/sutton/sutton-90.ps.gz | 9.541e-5 | 17(5), 2579(11), 573(17), 906(5), 638(9) |

Table 5: This table shows the top ranked pages according to the original Pagerank algorithm and the ranks achieved by these pages on the modified algorithm when run with 10, 15, 20,25 and 30 clusters respectively as shown in column 4. The value in the parenthesis indicate the block rank of the cluster to which the page belonged. One observation to make is the noisy nature of the block ranks. It appears that the clustering algorithm under the current parameters cannot identify good clusters consistently.

# References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," 1998.

[2] T. Haveliwala, "Topic-sensitive pagerank," in *Proceedings of the Eleventh International World Wide Web Conference, Honolulu, Hawaii*, 2002.

[3] S.Acharyya and J.Ghosh, "A maximum entropy framework for link analysis on directed graphs," in *LINKKDD'03 Workshop on Link Analysis for Detecting Complex Behavior, Washington D.C.*, 2003.

[4] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub, "Exploiting the block structure of the web for computing pagerank," tech. rep., Department of Computer Science, Stanford University, 2003.

[5] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pp. 89–98, 2003.

[6] S. Chien, C. Dwork, R. Kumar, and D. Sivakumar, "Towards exploiting link evolution," in *Workshop on Algorithms and Models for the Web Graph*, 2001.

[7] A. Y. Ng, A. X. Zheng, and M. I. Jordan, "Link analysis, eigenvectors and stability," in *IJCAI-01*, pp. 903–910, 2001.

[8] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.