Page Ranking Systems: Axiomatisation and Experimentation

HEPPENSTALL, ALAN

Thesis submitted for the degree of Master of Science by Research

Page Ranking Systems

Axiomatisation and Experimentation

Alan Heppenstall

2014

Durham University
Engineering and Computing Sciences

Contents

Statement of Copyright

Acknowledgments

First and foremost I would like to thank Stefan Dantchev, who has been my supervisor throughout my research. His constant assistance, patience and ability to guide me has been essential to my work.

I would like to thank Aidan Chalk for his regular assistance. Without his abilities to optimize code, grasp complex ideas and work through errors I would not have been able to construct this thesis. I would also like to thank Elizabeth Woodhouse for regularly proof reading and checking my work.

I would like to thank Alon Altman, Moshe Tennenholtz, Ignacio Palacios-Huerta, Oscar Voltij and Marc Najork for their work in ranking which my work extends.

I would like to thank the Department of Engineering and Computer Sciences for granting me the scholarship to undertake this research and my parents for their support during my studies.

Abstract

Ranking a set of objects based on the relationships between them is fundamental for use with search engines, e-commerce websites and in the field of bibliometrics. Two of the most prominent search ranking algorithms are PageRank and SALSA (Stochastic Approach to Link-Structure Analysis).

In this thesis, we further explore the connections between page ranking algorithms and the theory of social choice, providing a basis for theoretical assessment of a weighted version of PageRank and we create and assess a new page ranking algorithm, combining ideas from both PageRank and SALSA which we call Query-Independent SALSA.

We justify the use of weighted PageRank from a theoretical perspective by providing a set of axioms which characterize the algorithm. We provide a tighter bound for our derivation than that of Altman et al and show that each of our axioms are independent.

We describe a query-independent version of SALSA, using ideas from the PageRank algorithm and test this on a real-world subgraph of the web graph. We find that our new algorithm, Query-Independent Stochastic Approach to Link-Structure Analysis (QISALSA) slightly outperforms PageRank on two measures and under-performs on one measure. We suggest that the approach of combining aspects of both algorithms may be less effective than precomputational methods for query-dependent algorithms.

Abbreviations, Notations and Definitions

We provide a short list of less commonly used language and notation to assist the reader.

Symbol	Meaning	Source
$\mathcal{G}_V$	The set of all graphs for vertex set V	[2]
$S_G(v)$	The set of vertices in G which have an incoming edge from v	Page 10
$P_G(v)$	The set of vertices in G which have an incoming edge to v	Page 10
$\preceq$	An ordering	Page 10
Ranking System	A functional for every finite vertex set V maps $G \in \mathcal{G}_V$ to an ordering $\preceq^F_G \in L(V)$	Page 11
$\mathcal{W}(v_i, v_j)$	Weight preference of vertex v_i for v_j	Page 12
$\mathbf{W}_G$	A preference matrix of weighted preferences for each vertex in a graph G.	Page 20
$\mathbf{r}$	A rank vector for G derived from $\mathbf{W}$	Page 24
Inlinks	A ranking method that counts the number of incoming links to a page	Page 40
PageRank	A popular query-independent ranking method	Page 40
HITS	Hyperlink-Induced Topic Search - a popular query-dependent ranking method	Page 41
SALSA	Stochastic Approach for Link-Structure Analysis - a popular query-dependent ranking method	Page 41
MAP	Mean Average Precision - an effectiveness measure for ranking algorithms	Page 45
MRR	Mean Reciprocal Rank- an effectiveness measure for ranking algorithms	Page 45
NDCG	Normalized Discounted Cumulative Gain - an effectiveness measure for ranking algorithms	Page 46

1 Introduction

1.1 Web Mining

The World Wide Web is a huge collection of hyperlinked documents. It represents the largest, most democratic and most open publishing medium in the world. Web mining is defined as the application of data mining techniques, methodologies and models to the data, structure and usage of the World Wide Web. We can divide this up into three categories: web structure mining, web content mining and web usage mining [20]. Web structure mining involves mining the structure of the web graph whereby pages are represented by vertices and edges by hyperlinks between documents. Web content mining aims to discover useful information from web data, content and services. Web usage mining deals with secondary data such as server access logs generated by user interaction with the web [24].

1.2 Web Structure Mining

Web structure mining concerns the inter document structure of the World Wide Web via examining the hyperlinks between documents. The structure of the web can be viewed as a graph with pages represented by vertices and hyperlinks represented by edges. Structure mining reveals additional information about the documents in the relationships between them. For example we may view a hyperlink as a recommendation, analogous to a bibliographical citation. In this setting a large number of hyperlinks may indicate a popular document which in turn may indicate high quality content within that document [24].

1.3 Search Ranking

Modern web search engines crawl documents on the web and then build an index of meta data about these pages. A user enters a request to the search engine for documents relevant to their query. The engine then lists documents it deems most relevant to this term in descending order. Relevancy is commonly calculated using a hybrid method of web structure mining and content analysis based on the meta data in the engine's index [15].

For a given keyword or set of keywords, we may find the set of relevant pages to be extremely large [20]. However, a human user typically only looks at the first ten to twenty results [17]. Due to this abundance of information the problem of deciding which pages from the result set are most relevant is of key importance.

Ranking by popularity was first suggested in the late 1990s [14, 23]. Despite the lack of editorial review process on the web, there is a rich structure which can be utilized to estimate the relative popularity of pages. The initial algorithms proposed which popularized use of link structure as a measure of page popularity and therefore relevance were PageRank and HITS (Hypertext-Induced Topic Search).

These algorithms can generally be divided into those that depend on the query (query-dependent) and those in which a ranking/relevancy score can be pre-computed as it is independent of a user's query (query-independent).

1.4 Theoretical Search Algorithm Analysis

Whilst there is an extensive volume of research which focuses on creating, improving and experimenting with ranking systems there is only a limited body of work on the theoretical reasoning to support the use of one algorithm above another. We look to extend this knowledge by providing an axiomatisation of a weighted version of PageRank and justifying each of these axioms. This is one of the contributions of this thesis.

1.5 Improving Ranking Algorithms

Previous research has found that query-dependent algorithms excel at producing an effective ranking of documents when compared to query-independent methods [18, 21]. The main concern for use in real-world environments is the unacceptable delay in calculating such a score at query-time. Query-independent algorithms excel in performance for users as they can be precomputed and thus cause little delay when a user requests a ranking of documents. In our second contribution, we formulate an algorithm that aims to combine ideas from common algorithms of both types, PageRank and SALSA. This algorithm aims to combine the ability to precompute scores with effective ranking.

1.6 Theory of Social Choice

The theory of social choice combines individual preferences to reach a collective decision within a theoretical framework. Voting systems and therefore search ranking systems fit into this framework and abide by the same rules.

1.7 Thesis Structure

The remainder of this thesis is organised as follows: Chapter 2 contains an introduction to PageRank, Edge Weighted PageRank and improvements to prior work; Chapter 3 contains a Combinatorial Axiomatisation of Edge Weighted PageRank with proof of the axiomatisation; Chapter 4 includes the creation of a new search ranking algorithm Query-Independent Stochastic Approach to Link-Structure Analysis and a comparison of this to popular search ranking algorithms.

2 Theoretical evaluation of Search Ranking Algorithms

2.1 Introduction

Ranking a set of objects based on the relationships between them is fundamental for use with search engines, e-commerce websites and in the field of bibliometrics (see e.g. [10, 26]). Two famous and highly utilized examples are Google's PageRank [23] and eBay's reputation system [26]. Page ranking is most commonly associated with search engines, and in particular assists with the problem of information abundance. Often there exists a very large number of documents related to a particular query and the most relevant or important objects must be identified in a computationally efficient but effective manner [20].

An extensive volume of research has been created in the domain of page ranking (see [15, 20, 10]). Many of these focus on creating, improving and experimenting with ranking systems. Experimental surveys have been carried out using relevancy scores based on human expertise applied to subgraphs of the web graph in order to directly compare ranking systems [21]. However, we have only been able to find and strongly beleive there exists only a limited body of work (see [25, 2]) on the theoretical reasoning to support the use of one particular system above another.

The rest of this chapter is organized as follows: section 2.2 defines PageRank; Edge Weighted PageRank and states the notation we make use of; section 2.3 details a short amendment to the Altman et al paper and the final section summarises the chapter.

2.2 PageRank & Edge Weighted PageRank

As per Altman et al [2], we rank pages based on the stationary probability distribution of performing a random walk on a graph, where each vertex represents a page and each directed edge represents a hyperlink. This forms the basis for Google's PageRank. We restrict our attention to strongly connected graphs:

Definition 1. A directed graph is called *strongly connected* if for all vertices $v_1, v_2 \in V$ there exists a path from v_1 to v_2 in E [2].

Definition 2. Let $G = (V, E)$ be a directed graph, and let $v \in V$. The *successor set* of v is $S_G(v) = \{u|(v,u) \in E\}$, and the *predecessor set* of v is $P_G(v) = \{u|(u,v) \in E\}$ [2].

The output of a page ranking procedure can be viewed as an ordering of a set of options:

Definition 3. Let A be some set. A relation $R \subseteq A \times A$ is called an *ordering* on A if it is reflexive, transitive, complete and anti-symmetric. Let $L(A)$ denote the set of all possible orderings on A[2].

Remark 4. Let $\preceq$ be an ordering, then $\simeq$ is the equality predicate of $\preceq$. Formally, $a \simeq b$ if and only if $a \preceq b$ and $b \preceq a$ [2].

Given the above notation we can define what a ranking system is:

Definition 5. Let $\mathcal{G}_V$ be the set of all strongly connected graphs with vertex set V. A ranking system F is a functional that for every finite vertex set V maps every strongly connected graph $G \in \mathcal{G}_\mathrm{V}$ to an ordering $\preceq_F^G \in L(V)$ [2].

We define a *hyperlink matrix/adjacency matrix.* Search ranking algorithms commonly begin with such a structure as an input:

Definition 6. Let $G = (V, E)$ be a directed graph. The hyperlink matrix is defined as:

$$[H_G]_{i,j} = \begin{cases} 1 & if\ a\ hyperlink\ exists\ from\ page\ v_i\ to\ v_j \\ 0 & otherwise \end{cases}$$

We now define the PageRank matrix which captures the random walk created by the basic PageRank algorithm. Namely, in this process we start at a random page, and iteratively move to one of the pages that are linked to by the current page, assigning equal probabilities to each such page [2].

Definition 7. Let $G = (V, E)$ be a directed graph and assume that $V = \{v_1, v_2, ..., v_n\}$. The PageRank Matrix A_G (of dimension $n \times n$) is defined:

$$[A_G]_{i,j} = \begin{cases} 1/|S_G(v_j)| & (v_i, v_j) \in E \\ 0 & otherwise \end{cases}$$

The PageRank procedure will rank pages according to the stationary probability distribution obtained in the limit of the above random walk; this is formally defined as:

Definition 8. Let $G = (V, E)$ be some strongly connected graph and assume $V = (v_1, v_2, ..., v_n)$. Let $\mathbf{r}$ be the unique solution of the system $A_G \cdot \mathbf{r} = \mathbf{r}$ where $r_1 = 1$.

The PageRank $PR_G(v_i)$of a vertex $v_i \in V$ is defined as $PR_G(v_i) = r_i$. The PageRank ranking system is a ranking system that for the vertex set V maps G to $\preceq_G^{PR}$, where $\preceq_G^{PR}$ is defined as: for all $v_i, v_j \in V$: $v_i \preceq_G^{PR} v_j$ if and only if $PR_G(v_i) \leq PR_G(v_j)$ [2].

The above defines a powerful heuristic for ranking internet pages, as adopted by search engines [23]. We begin our contribution by modifying the above for strongly connected *weighted* graphs where the weight on each edge is proportional to its popularity.

We define a *weight function* for use in our Edge Weighted PageRank algorithm:

Definition 9. Let $G = (V, E)$ and $\mathcal{W}(u, v)$ be a function where (u, v) are a pair of vertices from V, then $\mathcal{W} : V \times V \to \mathbb{R}$.

Definition 10. Let $G = (V, E)$ be a strongly connected graph. We define a *weighted graph* as a strongly connected graph with a square matrix $W = \{w_{i,j}\}_{n \times n}$ with non-negative rational entries where each edge has a weight $w_{i,j}$, a measure of preference or vote of the page v_i for page v_j, as defined by $\mathcal{W}(u, v)$. Thus, a unit preference of the i-th page splits into fractions $^{w_{i,j}}/_{\sum_{j=1}^{n} w_{i,j}}$ among all other pages (including itself).

Our Edge Weighted PageRank still ranks pages according to the stationary probability distribution obtained in the random walk as per PageRank but the PageRank matrix now denotes the weights on edges:

Definition 11. Let $G = (V, E)$ be a directed graph and assume that $V = \{v_1, v_2, ..., v_n\}$. The Edge Weighted PageRank Matrix W_G (of dimension $n \times n$) is defined:

$$[W_G]_{i,j} = \begin{cases} 1/x & x = \mathcal{W}(v_i, v_j) \\ 0 & if\, \mathcal{W}(v_i, v_j) = 0 \end{cases}$$

The Edge Weighted PageRank procedure will rank pages according to the stationary probability distribution obtained in the limit of the above random walk, as per Definition 19. We aim to treat Edge Weighted PageRank from an axiomatic social choice perspective so in contrast to the numerical procedure we have just defined, we will provide a graph-theoretic, ordinal representation of Edge Weighted PageRank in chapter 3.

2.3 Altman Amendments

Before providing a graph-theoretic, ordinal representation of Edge Weighted PageRank we make two small modifications to the the work provided by Altman et al [2]: the removal of a redundant axiom and a more concise proof for one of the properties.

2.3.1 Self-edge and Isomorphism

We begin by restating the required self-edge and isomorphism axioms provided by Altman et al [2]:

Definition 12. (Isomorphism axiom) A ranking system F satisfies isomorphism if for every isomorphism function $\varphi : V_1 \longmapsto V_2$, and two isomorphic graphs $G \in \mathcal{G}_{V_1}, \varphi(G) \in \mathcal{G}_{V_2} : \preceq^F_{\varphi(G)} = \varphi(\preceq^F_G)$ [2].

Notation: Let $G = (V, E) \in \mathcal{G}_v$ be a graph such that $(v, v) \notin E$. Let $G' = (V, E \cup \{(v, v)\})$. Let us denote **SelfEdge**$(G, v) = G'$.

Definition 13. (Self-edge axiom) Let F be a ranking system. F satisfies the self-edge axiom if for every vertex set V and for every vertex $v \in V$ and for every graph $G = (V, E) \in \mathcal{G}_v$ such that $(v, v) \notin E$, and for every $v_1, v_2 \in V \backslash \{v\}$: Let $G' =$ **SelfEdge**(G, v). If $v_1 \preceq^F_G v$ then $v \not\preceq^F_{G'} v_1$; and $v_1 \preceq^F_G v_2$ iff $v_1 \preceq^F_{G'} v_2$ [2].

2.3.2 Vote by Committee Axiom

We show that the vote by committee axiom is redundant as it is equivelent to the combined use of two of the other axioms. We first re-state the axioms defined by Altman et al to be used in this section:

Definition 14. (Vote by committee) Let F be a ranking system. F satisfies vote by committee if for every vertex set V, for every vertex $v \in V$, for every graph $G = (V, E) \in \mathcal{G}_v$, for every $v_1, v_2 \in V$, and for every $m \in \mathbb{N}$: Let $G' = (V \cup \{u_1, u_2, ..., u_m\}, E \backslash \{(v, x) | \, x \in S_G(v)\} \cup \{(v, u_i) | \, i = 1, ..., m\} \cup \{(u_i, x) | \, x \in S_G(v), i = 1, ..., m\})$ where $\{u_1, u_2, ..., u_m\} \cap V = \emptyset$. Then $v_1 \preceq^F_G v_2$ iff $v_1 \preceq^F_{G'} v_2$ [2].

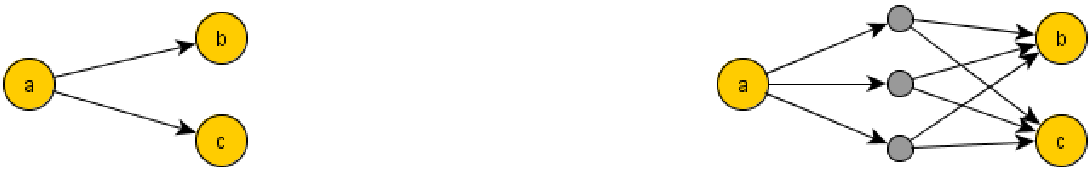


Figure 2.1: Vote by committee axiom

Definition 15. (Collapsing) Let F be a ranking system. F satisfies collapsing if for every vertex set V, for every $v, v' \in V$, for every $v_1, v_2 \in V \backslash \{v, v'\}$ for every graph $G = (V, E) \in \mathcal{G}_v$ for which $S_G(v) = S_G(v')$, $P_G(v) \cap P_G(v') = \emptyset$, and $[P_G(v) \cup P_G(v')] \cap \{v, v'\} = 0$: Let $G' = (V \backslash \{v'\}, E \backslash \{(v', x) | x \in S_G(v')\} \backslash \{(x, v') | x \in P_G(v')\} \cup \{(x, v) | x \in P_G(v')\})$. Then $v_1 \preceq^F_G v_2$ iff $v_1 \preceq^F_{G'} v_2$ [2].

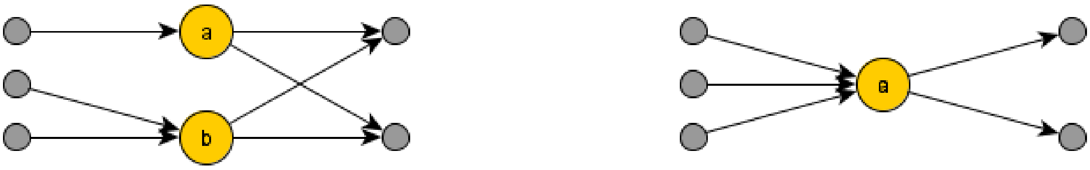


Figure 2.2: Collapsing axiom

Definition 16. (Proxy) Let F be a ranking system. F satisfies proxy if for every vertex set V, for every vertex $v \in V$, for every $v_1, v_2 \in V \backslash \{v\}$, and for every graph $G = (V, E) \in \mathcal{G}_v$ for which $|P_G(v)| = |S_G(v)|$, for all $p \in P_G(v)$: $S_G(p) = \{v\}$, and for all $p, p' \in P_G(v)$: $p \simeq_G^F p'$: Assume $P_G(v) = \{p_1, p_2, ..., p_m\}$ and $S_G(v) = \{s_1, s_2, ..., s_m\}$. Let $G' = (V \backslash \{v\}, E \backslash \{(x, v), (v, x) | x \in V\} \cup \{(p_i, s_i) | i \in \{1, 2, ..., m\}\})$. Then $v_1 \preceq_G^F v_2$ iff $v_1 \preceq_{G'}^F v_2$ [2].

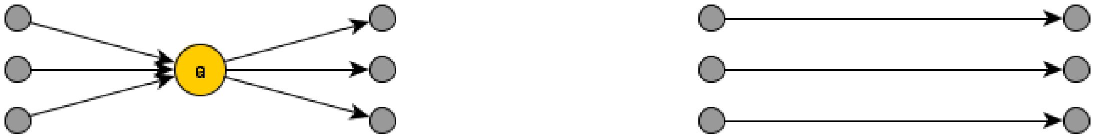


Figure 2.3: Proxy axiom

Lemma 17. *The Vote by committee axiom is equivalent to the combined use of the Collapsing axiom and Proxy axiom.*

Proof. We use the Proxy axiom in the reverse direction and then use Collapsing in the reverse direction, as illustrated in Figure 2.4

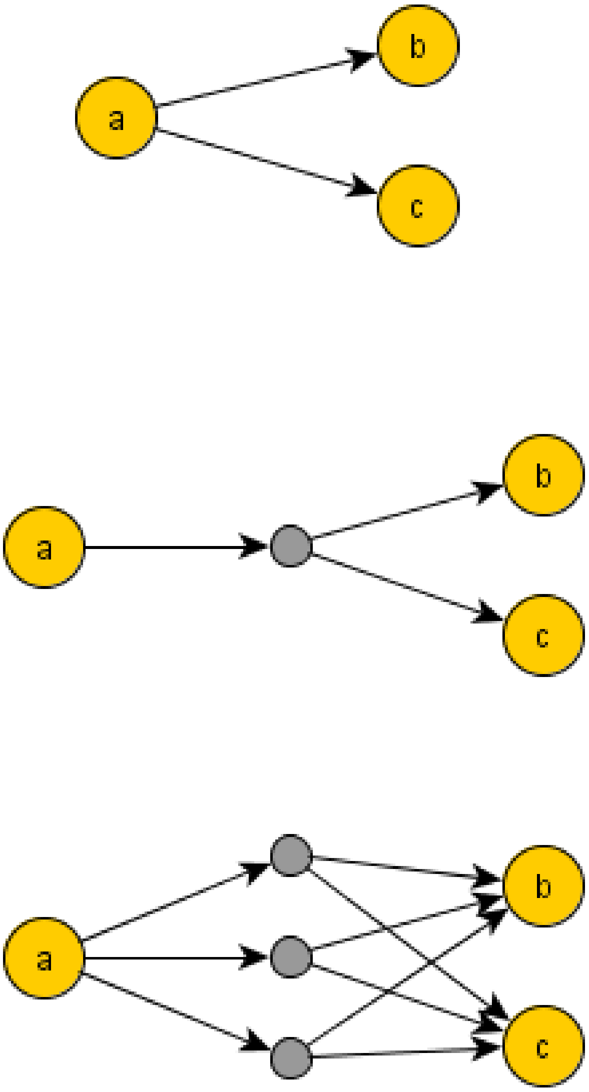


Figure 2.4: Illustration of proof of Lemma 4

□

2.3.3 Del Property

We provide a more concise proof for the Del property provided by [2] and begin by restating the required property definitions:

Definition 18. Let V be a vertex set and let $v \in V$ be a vertex. Let $G = (V, E) \in \mathcal{G}_v$ be a graph where $S(v) = \{s\}$, $P(v) = \{p\}$, and $(s, p) \notin E$. We will use $\mathbf{Del}(G, v)$ to denote the graph $G' = (V', E')$ defined by:

$V' = V \backslash \{v\}$

$E' = E \backslash \{(p, v), (v, s)\} \cup \{(p, s)\}$

[2]

Definition 19. Let F be a ranking system. F has the weak deletion property if for every vertex set V, for every vertex $v \in V$ and for all vertices $v_1, v_2 \in V \backslash \{v\}$, and for every graph $G = (V, E) \in \mathcal{G}_v$ such that $S(v) = \{s\}$, $P(v) = \{p\}$, and $(s, p) \notin E$: Let $G' = \mathbf{Del}(G, v)$. Then, $v_1 \preceq_G^F v_2$ iff $v_1 \preceq_{G'}^F v_2$ [2].

Lemma 20. *Let F be a ranking system that satisfies Isomorphism, Vote by committee and Proxy. Then, F has the weak deletion property [2].*

We can simplify the above lemma given by Altman et al to weaken the required axiom satisfaction to the following:

Lemma 21. *Let F be a ranking system that satisfies Isomorphism and Proxy. Then, F has the weak deletion property.*

Proof. Let V be a vertex set, let $v \epsilon V$; $v_1, v_2 \in V \backslash \{v\}$ be vertices and let $G = (V, E) \in \mathcal{G}_v$ be a graph such that $S(v) = \{s\}$, $P(v) = \{p\}$, and $(s, p) \notin E$. Assume that $v_1 \preceq_G^F v_2$. Let $s_0 = v$ and $S(p) = \{s_0, s_1, ..., s_m\}$. Let $G_1 = (V_1, E_1)$, where $V_1 = V \backslash \{v\}$, $E_1 = E \backslash \{(p, v), (v, s)\} \cup \{(p, s)\}$. By the Proxy axiom where $|S(v)| = 1$ and $|P(v)| = 1$, $v_1 \preceq_{G_1}^F v_2$. Let $G' = \mathbf{Del}(G, v)$. By the proxy and Isomorphism axioms $v_1 \preceq_{G'}^F v_2 \iff v_1 \preceq_{G_1}^F v_2$. Thus, $v_1 \preceq_{G'}^F v_2$ as required. □

2.4 Discussion

In this chapter we have provided the required definitions for our axiomization and have outlined the PageRank algorithm, the Edge Weighted PageRank algorithm and have made some small ammendments to the work of Altman et al. We are now ready to provide a combinatorial axiomatisation of Edge Weighted PageRank.

3 Combinatorial Axiomatisation of Edge Weighted PageRank

3.1 Introduction

We extend the work of Altman et al [2] and Palacios-Huerta et al [25] to provide a combinatorial axiomatisation of a weighted version of PageRank. We provide a theoretical basis for the use of Edge Weighted PageRank. Our contribution is to provide a set of axioms for Edge Weighted PageRank whose derivation is polynomially bound to the size of the input graph. This extends the work of Altman et al to provide a tighter bound than in their derivation of PageRank, to devise a set of axioms for a weighted environment and to show that the axioms are logically independent. We have furthered their exploration of the connections between page ranking algorithms and the mathematical theory of social choice. We extend the work of Palacios-Huerta et al in that our axioms provide an ordinal, graph-theoretic representation of the ranking system and examine these in the context of the World Wide Web.

As per Altman et al we will treat the internet as a graph, where pages/vertices are agents and the hyperlinks between pages act as votes of preference. In this case the problem of page ranking becomes the problem of aggregating rankings into a global ranking. In the classical theory of social choice, as defined by Arrow [4], a set of agents is called to rank a set of options. The unique aspect of ranking web pages is that the set of agents and options coincide. The transitive effects of voting then need to be considered as agents may directly influence their own ranking by adjusting their votes to other pages.

The rest of this chapter is organized as follows: section 3.2 details our axioms and provides some intuitive description of how they operate; section 3.3 details the proof of soundness for our axioms; section 3.4 details the proof of completeness for our axioms; section 3.5 provides justification for the independence of each axiom and the final section contains a discussion of the results presented.

3.2 Axioms

From a social choice perspective, we can view each page in the web graph as an agent, where this agent prefers the pages it links to above the pages it does not link to. Page ranking is therefore the same problem as finding a social aggregation rule. We identify a set of simple, graph-theoretic axioms which characterise and satisfy Edge Weighted PageRank and do not refer to numeric computations.

While all the axioms are of the form "if and only if" we will sometimes refer to the axiom in only one direction in the intuitive, descriptive explanation to ensure that these illustrations are kept simple (in all cases the intuition holds in both directions).

Let $V = \{v_1, \ldots, v_n\}$ be a set of web-pages with a preference-matrix $\mathbf{W} = \{w_{i,j}\}_{n \times n}$, and let $V' = \{v'_1, \ldots, v'_n\}$ be a set of different web-pages with a preference-matrix $\mathbf{W}' = \left\{w'_{i,j}\right\}_{n \times n}$.

Definition 22. *(Scaling axiom)* If for every $u_i \in S_G(v)$ and $u'_i \in S_{G'}(v')$, each edge weight $\mathcal{W}(v, u_i) = \alpha$ and $\mathcal{W}(v', u'_i) = c\alpha$, where c is a positive constant, then for every vertex $v \in V'$, $v'_i \preceq v'_j$ if and only if $v_i \preceq v_j$.

If we modify the weights of all outgoing edges of v proportionally, then the relative ranking in G is retained. The scaling axiom tells us that the absolute values applied to weights do not matter for our ranking, that they only indirectly represent the probability of arriving at the page in a random walk. We are only interested in the relative weights within the graph, so modifying local weights within the graph has no effect on the rest of the weights, due to the implicit normalisation that takes place in the Edge Weighted PageRank algorithm.

Definition 23. *(Isomorphism axiom)* Assume that $v_i \in V$ where $G = (V, E)$ and $v'_i \in V'$ where $G' = (V', E')$ where G and G' are isomorphic. If $v_i = v'_i$ for every $v_i \in V$ and each edge $\mathcal{W}(v, u_i) = \mathcal{W}(v', u'_i)$, then for every i, j $v'_i \preceq v'_j$ if and only if $v_i \preceq v_j$.

The isomorphism axiom tells us that the ranking procedure should be independent of the names given to pages. No particular page is singled out to have a special ranking and the other of input does not affect the ranking produced by Edge Weighted PageRank.

Definition 24. *(Self-Preference axiom)Assume that $G = (V, E)$ and $G' = (V', E')$. If $(v_i, v_j) = (v'_i, v'_j)$ for every $e \in E$ with the only exception (v_k, v_k) for some k where $\mathcal{W}(v_k, v_k) < \mathcal{W}(v'_k, v'_k)$, then*

1. for every $i, j \neq k$, $v'_i \preceq v'_j$ if and only if $v_i \preceq v_j$, and
2. for every $i \neq k$ such that $v_i \preceq v_k$, $v'_i \prec v'_k$.

The self-preference axioms tell us that if vertex v_i ranks at least as high as v_j in our graph G where v_i has no self edges, then if we add a self edge to G, v_i should be ranked higher than v_j and all the other vertices in the graph should retain their ranking. Essentially if a page adds a link to itself then its ranking should be at least as high as before this link was added.

For the next two axioms, we let $V' = \left\{v'_1, \ldots, v'_{n+1}\right\}$ be a set of $n+1$ web-pages with a preference-matrix $\mathbf{W}' = \left\{w'_{i,j}\right\}_{(n+1)\times(n+1)}$.

Definition 25. *(Equivalence axiom)* If

1. $S_{G'}(v'_n) = S_{G'}(v'_{n+1}) = S_G(v_n)$,
2. $\{(v'_n, v'_{n+1}), (v'_{n+1}, v'_n)\} \notin E'$ and $\mathcal{W}(v'_n, v'_n) = \mathcal{W}(v'_{n+1}, v'_{n+1})$,
3. $P_{G'}(v'_n) = P_G(v_n)$ and $P_{G'}(v'_{n+1})$ does not include $\{v'_n, v'_{n+1}\}$,
4. $\mathcal{W}(v'_1, v'_n) + \mathcal{W}(v'_1, v'_{n+1}) = \mathcal{W}(v_1, v_n)$,
5. For all i and j, $\mathcal{W}(v_i, v_j) = \mathcal{W}(v'_i, v'_j)$,

then for every $i, j \notin \{n, n+1\}$, $v'_i \preceq v'_j$ if and only if $v_i \preceq v_j$.

If we make a copy of v_n and divide the weight from v_1 to v_n and v_{n+1} then the relative ranking in the graph is unchanged. The equivalence axiom allows a vertex to separate its vote between two vertices as long as they share the same successor set and the weights are retained. An example sketch of the equivalence axiom is shown in Figure 2.5.

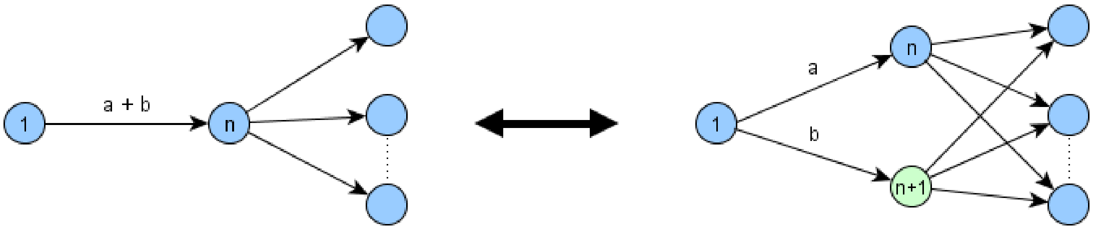


Figure 3.1: Equivalence axiom example

Definition 26. *(Proxy axiom)* If

1. $P_{G'}(v'_{n+1}) = \{v'_1\}$, $S_{G'}(v'_{n+1}) = S_{G'}(v'_1)$,
2. $\mathcal{W}(v'_1, v'_{n+1}) + \sum_{j \in S_{G'}(v_1)} \mathcal{W}(v'_1, j) = \sum_{j \in S_G(v_1)} \mathcal{W}(v_1, j)$ and
3. For all i and j, $\mathcal{W}(v_i, v_j) = \mathcal{W}(v'_i, v'_j)$,

then for every $i, j \neq n+1$, $v'_i \preceq v'_j$ if and only if $v_i \preceq v_j$.

If there is an additional vertex v_{n+1} between v_1 and a set of successors then we can remove v_{n+1} from the graph and retain the relative ranking for all other vertices. The proxy axiom essentially allows the creation or deletion of a 'dummy' page that redistributes a single vote to two other pages. An example sketch of the proxy axiom is shown in Figure 2.6.

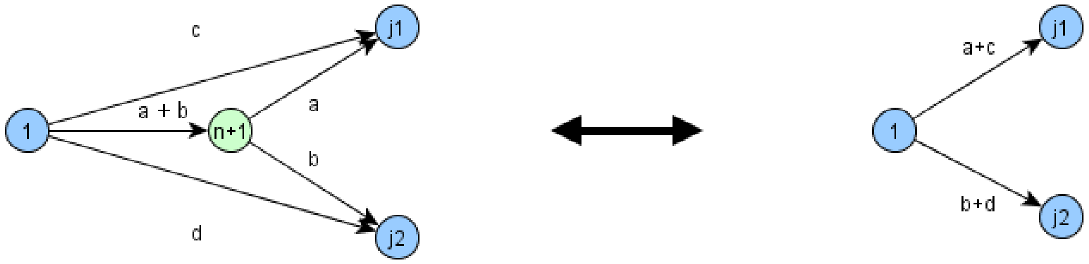


Figure 3.2: Proxy axiom example

We have provided some intuitive explanation of each axiom but one may argue that particular axiom(s) are not reasonable. However, we find that this set of axioms characterises Edge Weighted PageRank exactly and that all of these axioms are logically independent as shown in section 3.5.

3.3 Soundness

Proposition 27. *Edge Weighted PageRank satisfies the scaling, isomorphism, self preference, equivalence and proxy axioms.*

Remark 28. For all G, the weighted matrix for G is defined as:

$$\mathbf{W}_G = \begin{pmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \cdots & w_{n,n} \end{pmatrix}$$

and the rank vector for G is defined as:

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}$$

Proof. (Scaling) This axiom is satisfied directly by definition due to the normalisation in the algorithm.

(Isomorphism) This axiom is satisfied directly from the definition by the assumption that $V = \{v_1, v_2, ..., v_n\}$.

(Self preference) Let $V = \{v_1, v_2, \ldots, v_n\}$, and let $G = (V, E)$. Let $G' = (V', E')$ where $V' = V$ and $E' = E \cup \{(v_1, v_1, \alpha)\}$ where α is an edge weight. Let $\mathbf{r}$ be the solution of $\mathbf{W}_G \cdot \mathbf{r} = \mathbf{r}$, where $r_1 = 1$. Now let us show that the ranking remains the same in G and G'. For all $v \in V$:

$$[\mathbf{W}_G\mathbf{r}]_i = \sum_{j=1}^{n} w_{i,j} r_j = r_i$$

As the only element in $\mathbf{W}$ modified by the axiom is $w_{1,1}$ the above will hold for all cases in G' except $[\mathbf{W}_{G'}\mathbf{r}']_1$. So, we get $\mathbf{W}_{G'}\mathbf{r}' = \mathbf{r}'$ as required. For v_1:

$[\mathbf{W}_G\mathbf{r}]_1 = \sum_{j=1}^{n} w_{1,j}r_j$

but in G'

$[\mathbf{W}_{G'}\mathbf{r}']_1 = \alpha r'_1 w'_{1,1} + \sum_{j=1}^{n} w'_{1,j}r'_j$

where $\alpha = \frac{W'_{1,1}+\sum_{j=2}^{n} W_{i,j}}{W_{1,1}+\sum_{j=2}^{n} W_{i,j}}$

The weighted matrix for G' is:

$$\mathbf{W}_{G'} = \begin{pmatrix} \frac{\alpha}{n}w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ \frac{\alpha}{n}w_{n,1} & \cdots & w_{n,n} \end{pmatrix}$$

The rank vector for G' is:

$$\mathbf{r}' = \begin{pmatrix} \alpha r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$$

So, we get $[\mathbf{W}_{G'}\mathbf{r}']_1 = \mathbf{r}' + \alpha$ as required (as the axioms require v_1 to rank greater than or equal in G' to its rank in G).

(Equivalence) Let $V = \{v_1, v_2, \ldots, v_n\}$, and let $G = (V, E)$. Let $G' = (V', E')$ where $V' = V \cup \{v_{n+1}\}$ and $E' = E \cup \{(v_1, v_{n+1}), (v_{n+1}, y) | y \in S_G(v_n)\}$. Let $\mathbf{r}$ be the solution of $\mathbf{W}_G \cdot \mathbf{r} = \mathbf{r}$, where $r_1 = 1$. For all $v \in V$:

$$[\mathbf{W}_G\mathbf{r}]_i = \textstyle\sum_{j=1}^{n} w_{i,j} r_j = r_i$$

and for all $v' \in V' \backslash \{v'_n, v'_{n+1}, y | y \in S_G(v'_n)\}$:

$$[\mathbf{W}_{G'}\mathbf{r}]_i = \textstyle\sum_{j=1}^{n} w_{i,j} r_j = r_i$$

For v'_{n+1}:

$$[\mathbf{W}_{G'}\mathbf{r}]_{n+1} = w_{n+1,1} r_1$$

For v'_n:

$$[\mathbf{W}_{G'}\mathbf{r}]_n = [W_G r]_n - [W_{G'} r]_{n+1}$$

So ranking is retained between G and G' for all $v \in V \backslash \{v_n, S_G(v'_n)\}$.

For all $\{y | y \in S_G(v'_n)\}$ the ranking of y is:

$$[\mathbf{W}_{G'}\mathbf{r}]_y = w_{i,n} r_n + w_{i,n+1} r_{n+1} + \textstyle\sum_{j=2}^{n-1} w_{i,j} r_j = [\mathbf{W}_G\mathbf{r}]_y + w_{i,n+1} r_{n+1}$$

The weighted matrix for G' is:

$$\mathbf{W}_{G'} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & \cdots & \cdots & \frac{1}{2}w_{1,n} & w_{1,n+1} \\ w_{2,1} & w_{2,2} & \cdots & \cdots & \cdots & w_{2,n} & w_{2,n+1} \\ w_{3,1} & w_{3,2} & \cdots & \cdots & \cdots & w_{3,n} & w_{3,n+1} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \frac{1}{2}w_{n,1} & \frac{1}{2}w_{n,2} & \ddots & \ddots & \ddots & \frac{1}{2}w_{n,n} & \frac{1}{2}w_{n,n+1} \\ w_{n+1,1} & w_{n+1,2} & \cdots & \cdots & \cdots & w_{n+1,n} & w_{n+1,n+1} \end{pmatrix}$$

where $w_{i,n+1} = \frac{1}{2}w_{i,n}$. The rank vector for G' is:

$$\mathbf{r}' = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n - r_{n+1} \\ r_{n+1} \end{pmatrix}$$

In this case ranking is retained due to the scaling axiom. The addition of edges from v_{n+1} to each successor is equivalent to scaling the edges from v_n to each successor $S_G(v_n)$ by a constant factor given that $S_G(v_n) = S_{G'}(v_{n+1})$.

For v_n the ranking is:

$$[\mathbf{W}_{G'}\mathbf{r}]_n = w_{i,n}r_n - w_{i,n+1}r_{n+1} + \sum_{j=2}^{n-1} w_{i,j}r_j = [\mathbf{W}_G\mathbf{r}]_n - w_{i,n+1}r_{n+1}$$

and the weighted matrix for G' and rank vector remain as above. For v_n ranking is retained due to the scaling axiom. As $P_G(v_n) = P_{G'}(v_n)$ and $P_{G'}(v_n) = P_{G'}(v_{n+1})$ the addition of edges from $P_{G'}(v_n)$ to v_{n+1} is equivalent to scaling the edges from $P_{G'}(v_n)$ to v_n by a constant factor of $\frac{1}{2}$.

(Proxy) Let $V = \{v_1, v_2, \ldots, v_n\}$, and let $G = (V, E)$. Let $V' = V$ and $G' = (V', E')$ where $V' = V \backslash \{v_2\}$ and $E' = E \backslash \{(v_1, v_2), (v_2, y) | y \in S_G(v_2)\}$. Let $\mathbf{r}$ be the solution of $\mathbf{W}_G \cdot \mathbf{r} = \mathbf{r}$, where $r_1 = 1$. For all $v \in V \backslash \{v_3, v_4\}$:

$$[\mathbf{W}_G\mathbf{r}]_i = \textstyle\sum_{j=1}^{n} w_{i,j}r_j = r_i$$

For all $v' \in V' \backslash \{v'_3, v'_4\}$:

$$[\mathbf{W'}_{G'}\mathbf{r'}]_i = \textstyle\sum_{j=1}^{n-1} w'_{i,j}r'_j = r'_i$$

For v_3 and v_4:

$$[\mathbf{W}_G\mathbf{r}]_i = \textstyle\sum_{j=1}^{2} w_{i,j}r_j = \frac{w_{i,1}r_1}{|S_G(v_1) \cap S_G(v_2)|}$$

and for v'_3 and v'_4

$$[\mathbf{W'}_{G'}\mathbf{r'}]_i = w'_{i,1}r'_1 = \frac{w'_{i,1}r'_1}{|S_{G'}(v'_1)|}$$

The weighted matrix for G' is:

$$\mathbf{W}_{G'} = \begin{pmatrix} w_{1,1} & w_{2,3} + w_{1,3} & w_{2,4} + w_{1,4} & w_{1,5} & \cdots & w_{1,n-1} \\ w_{3,1} & w_{3,3} & \cdots & \cdots & \cdots & w_{3,n-1} \\ \vdots & \ddots & \cdots & \ddots & \ddots & \vdots \\ w_{n-1,1} & w_{n-1,3} & \cdots & \cdots & \cdots & w_{n-1,n-1} \end{pmatrix}$$

and the rank vector for G is:

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_{n-1} \end{pmatrix}$$

Since $|S_G(v_1) \cap S_G(v_2)| = |S_{G'}(v_1')|$ the rankings of v_3 and v_4 are retained as required.

□

3.4 Completeness

We now show that our axioms fully characterise the Edge Weighted PageRank system. We can prove:

Theorem 29. *A ranking system F satisfies scaling, isomorphism, self preference, equivalence and proxy if and only if F is the Edge Weighted PageRank ranking system.*

Given Theorem 29, it is enough to prove the following:

Proposition 30. *Let $F_1$and F_2 be ranking systems that satisfy scaling, isomorphism, self preference, equivalence and proxy. Then, F_1 and F_2 are the same ranking system.*

We shall now sketch the proof. Essentially we eliminate vertices, one after another whilst preserving the ranking of the other vertices. When we are left with the two vertices we would like to compare, we can compare the incoming weights of each vertex and decide which vertex has the largest incoming weight. The vertex with the larger incoming weight has the higher ranking. To do this effectively we must first equalise the weight between the two vertices and then compare the weight of the self preference edge.

The intuitive idea is to begin with our graph $G = (V, E)$ and two arbitrary vertices v_a and v_b in V and manipulate G by applying our axioms to achieve a new graph G_n for which F_1 and F_2 rank v_a and v_b the same as in G (formally $v_a \preceq^F_{G_n} v_b \iff v_a \preceq^F_G v_b$ for F_1 and F_2). One by one we remove vertices $v_i \neq \{v_a, v_b\}$ by replacing incident edges with direct edges from the predecessors of v_i to successors of v_i. Once we have a graph $G_m = (V_m, E_m)$ where $V_m = \{v_a, v_b\}$ we use scaling to ensure $\mathcal{W}(v_a, v_b) = \mathcal{W}(v_b, v_a)$. The relative ranking of v_a and v_b is retained throughout this process. We then use the self-edge axiom to create an isomorphic graph, showing that $v_a \preceq^F_{G_m} v_b \iff v_a \preceq^F_G v_b$. The steps required are:

1. We choose a vertex $v_x \neq \{v_a, v_b\}$ in the graph G. We replace v_x with an edge from each successor and predecessor of v_x while maintaining the relative ranking of v_a and v_b using the following steps:

   a) For each predecessor $P_G(v_x)$,

      i. Duplicate v_1 using the equivalence axiom where $v_x = v_n$ and $v_1 \in P_G(v_x)$.

   b) In the resulting graph G_{n+1}, for each new vertex v_{x_i},

      i. Use the proxy algorithm to remove v_{x_i} from G_{n+1}.

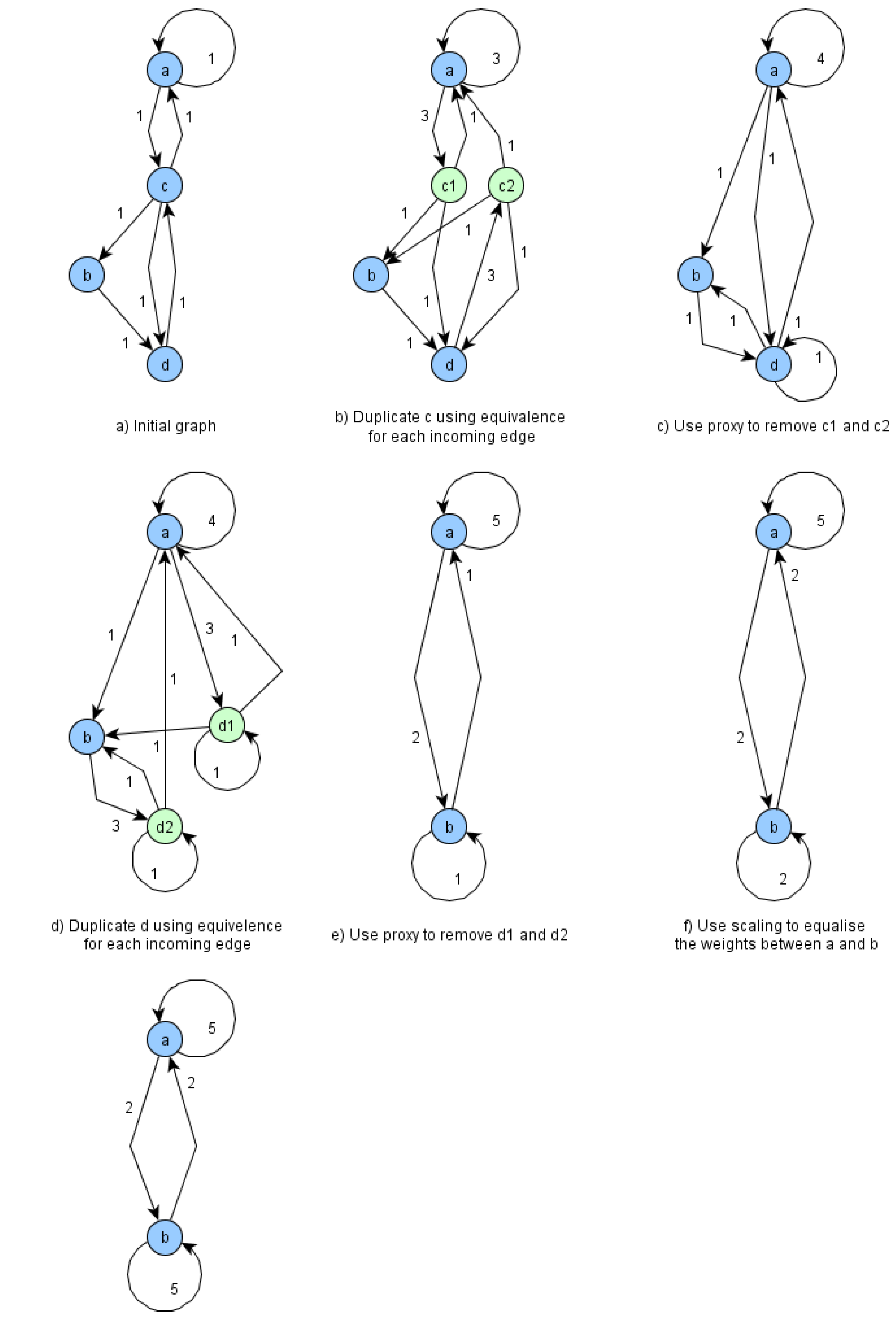
2. We repeat 1. until $G_m = (V_m, E_m)$ where $V_m = \{v_a, v_b\}$.

3. We use the scaling axiom to equalise $\mathcal{W}(v_a, v_b)$ and $\mathcal{W}(v_b, v_a)$ so that $\mathcal{W}(v_a, v_b) = \mathcal{W}(v_b, v_a)$.

4. We now add weight to the self edge of $v \in \{v_a, v_b\}$ using the self preference axiom so that the incoming and outgoing weights of v_a and v_b are equal. Let $v' = \{v_a, v_b\} \backslash \{v\}$. By the self edge axiom, if $v' \preceq^F v$ before adding the self edges, then now $v \not\preceq^F v'$ for F_1or F_2.

5. By the isomorphism axiom, in this graph $v_a \simeq v_b$, so before Step 4, $v' \preceq^F v$ for F_1 or F_2. However as the relative ranking of v_a and v_b did not change until Step 4, $v' \preceq_G^F v$ for $F \in \{F_1, F_2\}$ and thus $v_a \preceq_G^{F_1} v_b \Longleftrightarrow v_a \preceq_G^{F_2} v_b$.

Figure 3.3 shows an example of the completeness procedure where we rank v_a and v_b in G.



a) Initial graph

b) Duplicate c using equivalence for each incoming edge

c) Use proxy to remove c1 and c2

d) Duplicate d using equivelence for each incoming edge

e) Use proxy to remove d1 and d2

f) Use scaling to equalise the weights between a and b

g) Use self edge to make the graph isomorphic

Figure 3.3: Example Completeness Procedure

3.5 Independence of Axioms

We now show that our axioms are all logically independent using the method demonstrated by Palacios-Huerta et al [25]. We show that each axiom is independent to justify that all of our axioms are reasonable and that they are required as together a set to properly characterise Edge Weighted PageRank. Recall Definition 6 of a hyperlink matrix for a graph G and our description of a ranking system as per Definition 5. We now define a number of ranking methods, some of which are reasonable systems to order a set of pages while others are for demonstration purposes only. All are well defined and can produce an ordering from a connected weighted graph $G = (V, E)$.

Ranking systems:

1. *Egalitarian method.* This ranks every page equally by assigning each page the same rank. Formally, $F_E : \mathbb{R} \rightarrow \preceq^G_{F_E} \in L(V)$ is defined as $F_E(G) = (1/|V|, ..., 1/|V|)^T$ [25].
2. *Basic counting method.* This ranks each page based on the number of incoming links. Formally, $F_B : \mathbb{R} \rightarrow \preceq^G_{F_B} \in L(V)$ is defined as $F_B(G) = (\sum_{v \in V} h_{i,v}) i \in V$.
3. *Counting method.* This ranks each page based on the number of incoming links to the page divided by the total number of links in the graph. Formally, $F_C : \mathbb{R} \rightarrow \preceq^G_{F_C} \in L(V)$ is defined as $F_C(G) = (\frac{\sum_{v \in V} h_{i,v}}{\sum_{k \in V} \sum_{v \in V} h_{kv}}) i \in V$ [25].
4. *Invariant method.* This ranks each page according to the stationary distribution of the normalised adjacency matrix, similarly to a simplified PageRank method. Formally, $F_I : \mathbb{R} \rightarrow \preceq^G_{F_I} \in L(V)$ is defined as $F_I(G)$ where $F_I(G)$ returns the stationary distribution of the normalised adjacency matrix for the graph G as a vector [25].

5. *Out counting method.* This ranks each page according to the number of outgoing links from the page but does not allow self-references to affect the ranking. Formally $F_O : \mathbb{R} \rightarrow \preceq^G_{F_O} \in L(V)$ is defined as $F_O(G) = (\sum_{v \in V} h_{v,i}) i \in V$ and $v \neq i$.
6. *Normalised self reference method.* This ranks each page by normalising the weights across the graph so that the total outgoing weight of edges connected each vertex v_i sum to 1 and then counting the number of self-links. Formally $F_S : \mathbb{R} \rightarrow \preceq^G_{F_S} \in L(V)$ is defined as $F_S(G) = (\sum_{v \in V} h_{v,v}) v \in V$.

We show that our axioms are logically independent.

To see that the scaling axiom is independent of our other axioms, we consider the Out counting method, F_O. We can see that F_O satisfies isomorphism by definition. Self preference is satisfied as we do not count self edges as outgoing edges in our definition of F_O. Equivalence and proxy are satisfied as the weights are divided so that the sum of the new weight is equal to the sum of the original when duplicating or removing edges. Out counting does not satisfy scaling as we can increase the weight of outgoing edges for a particular vertex in G' using this axiom so the relative ranking is different in G and G'.

To see that the isomorphism axiom is independent of our other axioms, we consider a ranking system equivalent to the Invariant method F_I , except for some page $v_x \in V$ as the vector h_x we set $h_x = 1$. We can see that this new ranking method F_X satisfies all of our other axioms as it is equivalent to our Edge Weighted PageRank. This method F_X does not satisfy isomorphism as the ranking depends upon the naming of the pages. Therefore isomorphism is independent of the other axioms.

To see that self preference is independent of our other axioms, we consider the Egalitarian method, F_E. We can see that F_E satisfies isomorphism by definition. Equivalence and proxy are satisfied as F_E will maintain the preference in G and G'

despite the modifications to the graph. This method does not satisfy self preference as with the addition of edges to a vertex v'_k, $v'_i \preceq v'_k$ in G' if $v_i \preceq v_k$ in G but it should be $v_i \prec v_k$. Therefore self preference is the only one of our axioms not satisfied by F_E and so is logically independent.

To see that the equivalence axiom is independent of our other axioms, we consider the counting method F_C. We can see that F_C satisfies isomorphism by definition. Scaling is satisfied because we normalise the weights before taking the ranking, so local scaling cannot affect the relative ranking. Proxy is satisfied as the weights are divided equally so the incoming weights will remain the same in G and G'. Equivalence is not satisfied because we normalise then count; so the successor set of the vertices involved in the transformation have greater incoming weight in G' than in G.

To see that the proxy axiom is independent of our other axioms, we consider the the Normalised self reference method F_S. We can see that F_S satisfies isomorphism by definition and trivially satisfies self preference. Scaling is satisfied due to the normalisation that takes place initially. Equivalence is satisfied as all self links will be retained within the vertices involved in the transformation. Proxy is not satisfied as can be shown in a simple case where G has vertices $v_1, v_2 \in V$ and edges $(v_1, v_1, a), (v_1, v_2, b) \in E$ where a and b are the edge weights and in G' we have $v'_1, v'_2, v'_{n+1} \in V$ and the edges $(v'_1, v'_{n+1}, a+b), (v'_{n+1}, v'_1, a), (v'_{n+1}, v'_2, b)$ as allowed by the proxy axiom we find that we can introduce or remove a self edge and therefore modify the ranking of vertex v_1.

As all of our axioms are logically independent we are confident that no axiom can be simulated using another and that we require all of them to properly characterise the Edge Weighted PageRank ranking system.

3.6 Discussion

Representation theorems are the formal mathematical tool for the justification of decision and choice rules. By providing an ordinal, graph-theoretic representation of Edge Weighted PageRank we feel that a greater justification exists for use of PageRank in a weighted environment and that a basis has been created for further work in which pages are viewed as agents attempting to maximise their own utility. We have furthered the work by Altman et al to provide an axiomatisation of PageRank in which the derivation is polynomially bound by the size of the input graph.

It would be interesting to provide an axiomatisation for another ranking procedure such as Hubs and Authorities or the Stochastic Approach for Link-Structure Analysis. We feel that this axiomatisation may be simplified by examining each ranking procedure in a weighted environment. This would allow for a rigorous comparison and evaluation of ranking methods from a theoretical perspective.

Like Altman et al, we believe that the problem of ranking Internet pages is a fundamental problem that is intriguing and contains a varied array of open problems.

4 Query-Independent Stochastic Approach to Link-Structure Analysis

4.1 Introduction

Our motivation is to provide a link-structure analysis ranking algorithm which combines aspects of the Stochastic Approach to Link-Structure Analysis (SALSA) such as the concept of the two-step Markov Chain with the computational efficiency of a query-independent algorithm like PageRank.

We create a new algorithm, Query-Independent SALSA and test its performance according to various measures against PageRank and Inlinks.

The rest of this section is organised as follows: section 4.2 outlines the previous related work and specifies the original SALSA algorithm, 4.3 describes the data sets we have chosen for experimentation and how we evaluate algorithms, 4.4 defines the effectiveness measures used to compare ranking algorithms, 4.5 specifies the new QISALSA algorithm, 4.6 details our experiments and their results and 4.7 contains a discussion of the research.

4.2 Related Work

The idea of using link-structure analysis for ranking web pages in search results arose around 1997 and resulted in the creation of the HITS [14] and PageRank [23] algorithms. These algorithms have since become hugely popular and have spawned a large amount of related research, especially due to the commercial success of the Google search engine.

There have been numerous attempts at improving the effectiveness of HITS and PageRank in both quality of results and their computational efficiency. Query-independent algorithms inspired by PageRank include Topic-Sensitive PageRank [11], BlockRank [13], PowerRank [29], and PopRank [30]. Query-dependent algorithms inspired by HITS include PHITS [9], Randomised HITS [3] and most notably SALSA [16]. These attempts commonly aim to exploit an additional observation about the structure of the web graph in order to refine the given ranking but rarely change the fundamental thesis underlying PageRank or HITS.

A key difference between the PageRank and HITS authority scores are that PageRank is query-independent and thus can be computed off-line for the entire web graph whilst the HITS algorithm requires the construction of a neighbourhood graph based on the web pages that are related to the query and thus requires more computation at query-time [14, 23]. SALSA attempted to improve the effectiveness of both by combining features from HITS and PageRank but remains query-dependent [16].

More recently there has been increased research related to improving the computational performance of HITS and SALSA in the setting of a large-scale, real world implementation. It has been suggested that using Bloom Filters to precompute neighbourhoods of web pages may speed up HITS-like ranking algorithms at query-time [28]. Precomputing SALSA maps has been suggested to improve the algorithms efficiency and effectiveness. This approach computes a "score-map" for each web page in the web graph by performing a SALSA-like algorithm on its neighbourhood and retaining the scores of the most promising vertices in the neighbourhood graph [22]. Another approach to improve SALSA and similar algorithms is to use consistent sampling of vertices when constructing the neighbourhood graph and to use a reduced graph with fewer neighbours of relevant pages. This approach has been shown to perform efficiently and effectively in an experimental environment [19].

Relatively few evaluations of web page ranking algorithms have been produced in comparison to the volume of research related to improving their effectiveness, especially in a large-scale setting. There exists a number of small-scale studies such as that by Amento et al. [5] who employ quantitative measures but base their results on a set of just 5 queries and a graph induced by topical crawls related to the query. Borodin et al. [1] base their results on 34 queries, result sets of 200 pages per query and graphs derived from the first 50 in-links per result from the Google search engine. The first large-scale evaluation of HITS in comparison to other link-based ranking algorithms that we are aware of was performed by Najork et al. [18] and uses a graph covering 2.9 billion URLs and 28,000 queries. They found that HITS outperforms PageRank but is about as effective as measuring in-degree (the number of in-links). The only large-scale evaluation of SALSA in comparison to other link-based ranking algorithms that we are aware of was performed by Najork [21]. It uses a web graph induced by 463 million crawled web pages and uses 28,000 queries which includes 485,656 results labelled by human judges. This study found that SALSA substantially outperforms HITS.

4.2.1 Ranking Algorithms

To provide a base of comparison we will compare our new algorithm query-independent SALSA to PageRank and inlinks. Inlinks provides an established base measure and PageRank is the seminal algorithm of comparable complexity and design. We choose not to compare with the original SALSA as the query-dependent computation is not realistic in a real world setting without the addition of some type of pre-computation of node adjacency, something that is beyond the scope of our research. More generally we could formulate a query-independent version of most algorithms but with the limited scope possible within our experimental set up we prioritise comparison against similar query-independent algorithms. Here we outline inlinks, PageRank and SALSA. A description of HITS is also included to provide the context upon which SALSA is based.

4.2.1.1 Inlinks

To rank pages based on the number of inlinks to a vertex we simply count the number of incoming links for a page:

1. $G = (V, E)$ where G is the Web graph
2. Define $\mathbf{L}$ as the adjacency matrix for the graph G where
$$\mathbf{L}_{i,j} = \begin{cases} 1 & if\, an\, edge\, exists\, from\, L_i\, to\, L_j \\ 0 & otherwise \end{cases}$$
3. Compute ranking score of each vertex $v_i = \sum_{j_0}^{j_n} \mathbf{L}_{i,j}$

4.2.1.2 PageRank

We use a relatively simple implementation of PageRank from [15] with a standard α of 0.85:

1. $G = (V, E)$ where G is the Web graph
2. Define $\mathbf{P}$ as the adjacency matrix for the graph G where
$$\mathbf{P}_{i,j} = \begin{cases} 1/|P_i| & if\, an\, edge\, exists\, from\, P_i\, to\, P_j \\ 0 & otherwise \end{cases}$$
3. $\bar{\mathbf{P}} = \mathbf{P}$ where all rows consisting of only 0 are replaced with $\mathbf{e}^T/n$ where n is the order of $\mathbf{P}$ and $\mathbf{e}$ is a column vector of all ones.
4. $\alpha = 0.85$
5. Compute the ranking vector as follows:

   a) $\pi^{(k+1)T} = \alpha\pi^{(k)T}\bar{\mathbf{P}} + (1-\alpha)\mathbf{v}^T$

4.2.1.3 Hypertext-Induced Topic Search (HITS)

The thesis underlying HITS is that good authorities are pointed to by good hubs and good hubs point to good authorities. We calculate HITS as per [15] as follows:

1. $G = (V, E)$ where G is the Web graph
2. Define $\mathbf{L}$ as the adjacency matrix for the graph G where
$$\mathbf{L}_{i,j} = \begin{cases} 1 & if\, an\, edge\, exists\, from\, L_i\, to\, L_j \\ 0 & otherwise \end{cases}$$
3. Initialize $\mathbf{y}^0 = \mathbf{e}$ where $\mathbf{e}$ is a column vector of all ones.
4. Compute authority score as $\mathbf{x}_i^k = \mathbf{L^T y}^{(k-1)}$
5. Compute hub score as $\mathbf{y}_i^k = \mathbf{Lx}^{(k)}$
6. Compute $k = k + 1$ and normalize $\mathbf{x}^k$ and $\mathbf{y}^k$
7. Repeat steps 4 to 6 until convergence. [14]

4.2.1.4 Stochastic Approach for Link-Structure Analysis (SALSA)

The stochastic approach for link-structure analysis combines aspects of PageRank and HITS. It calculates a hub and authority score for each page in a manner similar to HITS but derives these from Markov Chains as per PageRank. The original SALSA algorithm formulated as per [15] is as follows:

1. Given a query q we build a graph $G = (V, E)$ induced from V where $V = \{v_i, \ldots, v_n, v_j, \ldots, v_m\}$; $\{v_i, \ldots, v_n\}$ is the set of pages from the Web graph which are directly relevant to q and $\{v_j, \ldots, v_m\}$ are the neighbours of and the neighbours of $\{v_i, \ldots, v_n\}$

2. Given the input graph $G = (V, E)$ create two sets V_h and V_a where $V_h = \{v_i | deg_{out}(v_i) > 0\}$ and $V_a = \{v_i | deg_{in}(v_i) > 0\}$

3. Create a bipartite graph $G' = (V', E')$ where $V' = V_h \cup V_a$ and $E' = \{(k_i, l_i) | k_i \epsilon V_h, l_i \epsilon V_a\}$

4. Define $\mathbf{L}$ as the adjacency matrix for the graph G' where
$$\mathbf{L}_{i,j} = \begin{cases} 1 & if\, an\, edge\, exists\, from\, i\, to\, j \\ 0 & otherwise \end{cases}$$

5. Define $\mathbf{L}_r$ as $\mathbf{L}$ with each nonzero row divided by its row sum and $\mathbf{L}_c$ be $\mathbf{L}$ with each nonzero column divided by its column sum

6. The hub matrix $\mathbf{H}$ consists of the nonzero rows and columns of $\mathbf{L}_r\mathbf{L}_c^T$ and the authority matrix $\mathbf{A}$ consists of the nonzero rows and columns of $\mathbf{L}_c^T\mathbf{L}_r$

7. Compute the hub and authority vectors as follows:

   a) $\pi_a^{(k+1)T} = \mathbf{L}_r^T\mathbf{L}_c\pi_a^{(k)T}$

   b) $\pi_h^{(k+1)T} = \mathbf{L}_c\mathbf{L}_r^T\pi_h^{(k)T}$

Remark 31. If G' is connected then these matrices are both irreducible Markov Chains and π_h^T, the stationary vector of $\mathbf{H}$, gives the hub scores for the query and

π_a^T, the stationary vector of $\mathbf{A}$, gives the authority scores for the query. If G' is not connected then $\mathbf{H}$ and $\mathbf{A}$ contain multiple irreducible components which must be calculated and then combined together to form the global ranking [16, 15].

Experimentally SALSA has shown to be more effective than PageRank or HITS [21]. Implementing SALSA for a large-scale search engine presents an issue common to all query-dependent link-structure ranking algorithms in that computation at query-time is too slow. The majority of this time is spent computing the neighbourhood graph related to a query. It has been shown that delaying the response time from a search engine by even a small measure leads to a significant drop in usage [17]. Research has been conducted, which aims to precompute aspects of the SALSA computation, which has shown to provide effective results which are derived more efficiently than SALSA but none of these have proven nearly as time efficient as a query-independent algorithm such as PageRank [21, 22, 19].

4.3 Design of Empirical Evaluation

Our experiments are based on a subgraph of the web graph and a set of queries with associated results, some of which are assessed for relevance using human expertise.

We make use of the TREC (Text REtrieval Conference) Category B subset of the ClueWeb09 Dataset. This subgraph of the web graph contains the first 50 million English pages from their web crawl with 428,136,613 unique URLs and a total of 454,075,638 outlinks. This dataset was crawled from the Web during January and February of 2009 [6] and provides a large subgraph of the Web Graph which we may be confident in using to assess our algorithm in an empirical manner.

The query set we use to measure relevance is the set of queries and graded results provided for the TREC 2010 Web Track and the TREC 2011 Web Track. Together these consist of 98 queries and a set of documents from the ClueWeb09 dataset which have been assessed for relevance using human expertise. They are graded as follows: -2 for spam or otherwise seems useless for any information need, 0 for not relevant, 1 for relevant, 2 for a page or site that is comprehensive and should be a top search result, 3 for a navigational result for the query [8, 7].

To empirically evaluate our algorithm we run it alongside inlinks and PageRank upon the TREC Category B set of web pages. We then construct a set of ranked results for each of the 98 queries from the ClueWeb09 dataset using the ranking scores provided by each algorithm. To properly compare the algorithms we use three effectiveness measures: Mean Average Precision, Mean Reciprocal Rank and Normalized Discounted Cumulative Gain. These effectiveness measures provide an assessment of each algorithm for each of the 98 keywords and the ranked results produced.

4.4 Effectiveness Measures

As per Najork [21], we use three performance measures to compare the effectiveness of each algorithm: mean average precision, mean reciprocal rank and normalised discounted cumulative gain. Given a rank-ordered vector of n results, let $rat(i)$ be the rating of the results at rank i, with 0 being detrimental, 1 being not relevant, 2 being relevant, 3 being very relevant and 4 being essential (we map these to the human graded results to compute effectiveness measures). Let $rel(i)$ be 1 if the result at rank i is relevant and 0 otherwise (we consider a result to be relevant if it has a label of "good" or better and irrelevant if it has a label of "fair" or worse). For all measures we will assume a document cut-off value k of 10 as studies have indicated that users commonly view only the top 10 results when performing a query with a search engine [27].

4.4.1 Mean Average Precision

Precision is a measure of how many of the retreived results were relevant to the query. The precision $P@k$ at document cut-off value k is defined as the fraction of relevant results among the k highest-ranking results:

$$\frac{1}{k}\sum_{i=1}^{k} rel(i)$$

Average precision considers the order in which the retreived documents are presented, meaning that a higher ranking relevant result is more desireable than a lower ranking relevant result. The average precision at cut-off value k is defined as:

$$AP@k = \frac{\sum_{i=1}^{k} rel(i) \cdot P@i}{\sum_{i=1}^{n} rel(i)}$$

where n is the total number of documents in the collection and thus the denominator is the total number of relevant results in the collection. The *mean average precision* $MAP@k$ at document cut-off value k of a query set is the mean of the average precisions of all queries in a query set.

4.4.2 Mean Reciprocal Rank

Reciprocal rank is the multiplicative inverse of the rank of the first relevant result in the top k. It provides a measure of quality of the order in which results are presented. A list of results correctly ordered by relevancy would give a score of 1 and the opposite 0. The reciprocal rank at document cut-off value k is defined as:

$$RR@k = \begin{cases} \frac{1}{i} & if\ \exists i \leq k : rel(i) = 1 \wedge \forall j < i : rel(j) = 0 \\ 0 & otherwise \end{cases}$$

The *mean reciprocal rank* $MRR@k$ is the mean of the reciprocal ranks of all queries in a query set at document cut-off value k.

4.4.3 Normalised Discounted Cumulative Gain

Discounted cumulative gain is based on the assumptions that highly relevant documents are more useful when they appear earlier in the search engine results and that the more relevant a document, the more useful it is. It is based upon the more primitive Cumulative Gain measure which measures the relevancy of documents returned; the sum of the graded relevance of the results for a query:

$$CG@k = \sum_{i=1}^{k} rat(i)$$

Discounted cumulative gain adds position in the result list into account with the assumption that earlier results should be those of higher relevance. We define the discounted cumulative gain at cut-off value k to be:

$$DCG@k = \sum_{i=1}^{k} \frac{(2^{rat(i)}-1)}{log_2(1+i)}$$

Normalized Discounted Cumulative Gain takes this measure and applies it accross a set of queries and result sets. We define the normalised discounted cumulative gain of a scored result set to be the given discounted cumulative gain divided by the 'ideal' discounted cumulative gain provided by an optimal scoring function:

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

In practice, the ideal discounted cumulative gain is created using the relevancy scores assessed using human expertise and thus is limited by their correctness.

4.4.4 Summary

MAP and MRR are simpler measures that provide less accuracy because they assume a document is releveant or not relevant; a simplified perspective of search users. NDCG provides a more detailed measure as it weights pages based on their relevance but is more reliant upon the subjective human expertise used to produce these weights.

We choose these measures to assess the effectiveness of QISALSA as they are the most commonly used in evaluating search ranking algorithms in more recent literature [12, 18, 21, 19, 22]. They provide a broad assessment of results set relevance and order.

4.5 Query-Independent SALSA

The modified, query-independent SALSA algorithm we create is defined as follows:

1. $G = (V, E)$ where G is the Web graph
2. Given the input graph $G = (V, E)$ create two sets V_h and V_a where $V_h = \{v_i | deg_{out}(v_i) > 0\}$ and $V_a = \{v_i | deg_{in}(v_i) > 0\}$
3. Create a bipartite graph $G' = (V', E')$ where $V' = V_h \cup V_a$ and $E' = \{(k_i, l_i) | k_i \epsilon V_h, l_i \epsilon V_a\}$
4. Define $\mathbf{L}$ as the adjacency matrix for the graph G' where
$$\mathbf{L}_{i,j} = \begin{cases} 1 & if\, an\, edge\, exists\, from\, i\, to\, j \\ 0 & otherwise \end{cases}$$
5. Define $\mathbf{L}_r$ as $\mathbf{L}$ with each nonzero row divided by its row sum and $\mathbf{L}_c$ be $\mathbf{L}$ with each nonzero column divided by its column sum
6. The hub matrix $\mathbf{H}$ consists of the nonzero rows and columns of $\mathbf{L}_r \mathbf{L}_c^T$ and the authority matrix $\mathbf{A}$ consists of the nonzero rows and columns of $\mathbf{L}_c^T \mathbf{L}_r$

7. Compute the hub and authority vectors as follows:

   a) $\pi_a^{(k+1)T} = (1-\varepsilon)\mathbf{L}_r^T\mathbf{L}_c\pi_a^{(k)T} + \frac{\varepsilon}{n}\mathbf{1}^T$

   b) $\pi_h^{(k+1)T} = (1-\varepsilon)\mathbf{L}_c\mathbf{L}_r^T\pi_h^{(k)T} + \frac{\varepsilon}{n}\mathbf{1}^T$

This algorithm pre-computes the scores and then simply performs a look-up at query time for the pages relevant to the query in a similar manner to PageRank. To assist with convergence and ensure that the vectors produced by QISALSA are unique we begin by modifying the normalized adjacency matrices of SALSA to ensure that they are irreducible. To achieve this we use the idea of the random surfer getting bored of their current state in the authority or hub side Markov chain and jumping to a random page with probability ε. This allows the computation to be performed on the entire web graph and therefore be precomputed. We add the identity matrix multiplied by the factor ε divided by the number of vertices in the graph to ensure that the graph becomes connected, is therefore irreducible and our computation will converge.

In some respect QISALSA is more akin to PageRank than SALSA but we believe it to be a valid and novel approach to ranking pages based on link-structure and merits experimentation to measure its efficiency and effectiveness.

4.6 Experimental Setup & Results

Experiments were conducted on two machines. The rankings were primarily calculated using a dual-core Intel Pentium CPU running at 2.8GHz with 4GB of memory and the calculations to grade effectiveness were primarily performed using a machine with 47 AMD Opteron CPUs running at 2.3GHz with 64GB or memory. We chose to calculate query-independent SALSA on a less powerful machine than available as a major motivation behind the creation of the algorithm was to ensure that performance could match other query-independent algorithms. All algorithms were written in Java and were executed within a Java Virtual Machine.

We performed ranking calculations for query-independent SALSA, PageRank and inlinks using the set of 98 queries on the document corpus. For each query we generated 1000 relevant results using a text-based search engine and then ordered these results using the chosen algorithm. A summary of the results with the mean values for each performance measure is shown in figure 4.1.

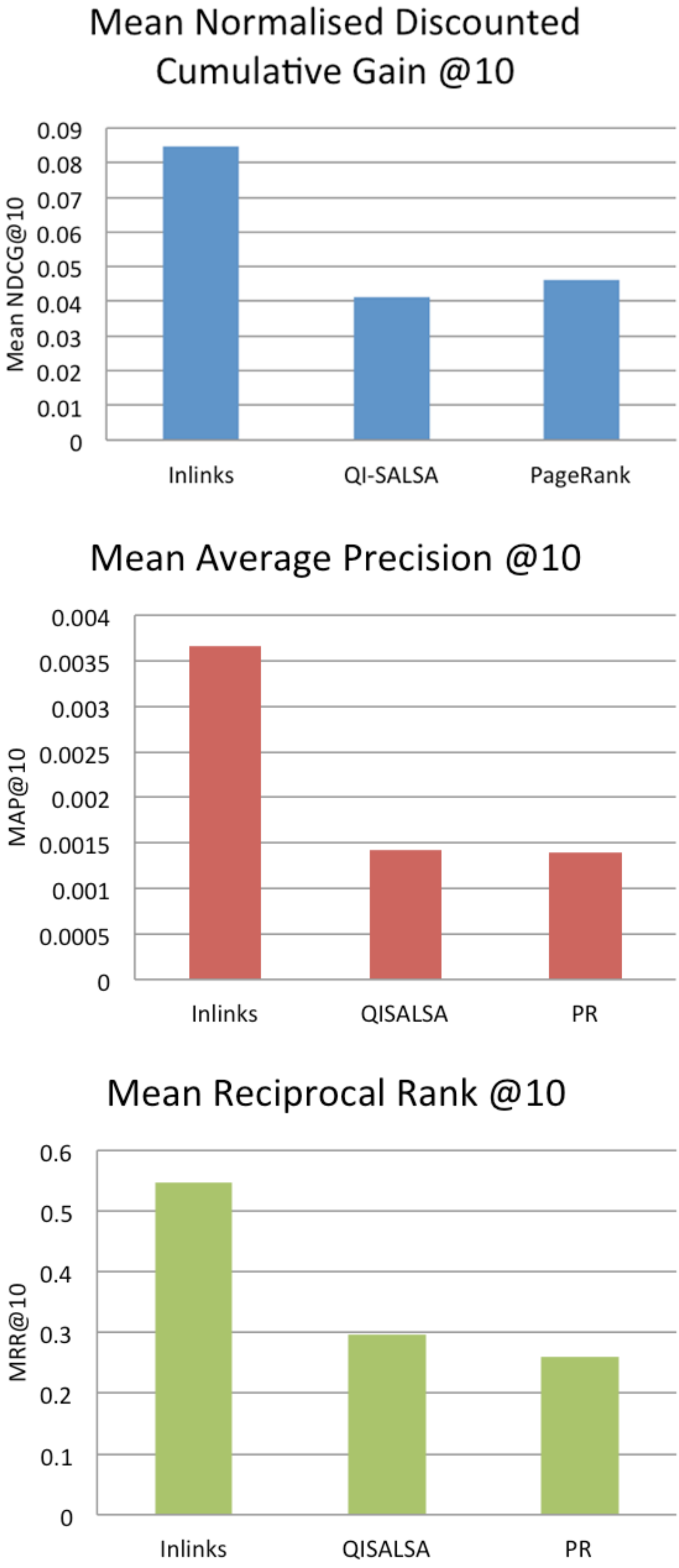


Figure 4.1: Summary of experimental results

The normalized discounted cumulative gain scores for each query for each algorithm are shown for each query in figure 4.2. The variance of these is shown in figure 4.3 which plots a linear trendline for each algorithm accross the resultset. As per previous results [18], on average inlinks performs better than PageRank. We find that QISALSA performs slightly worse than PageRank for this measure.

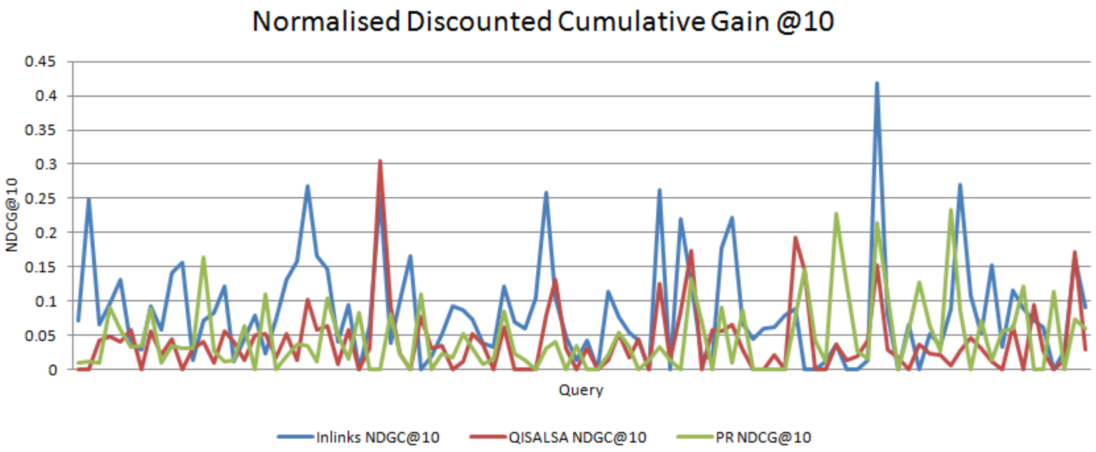


Figure 4.2: Normalized discounted cumulative gain @10 for each query

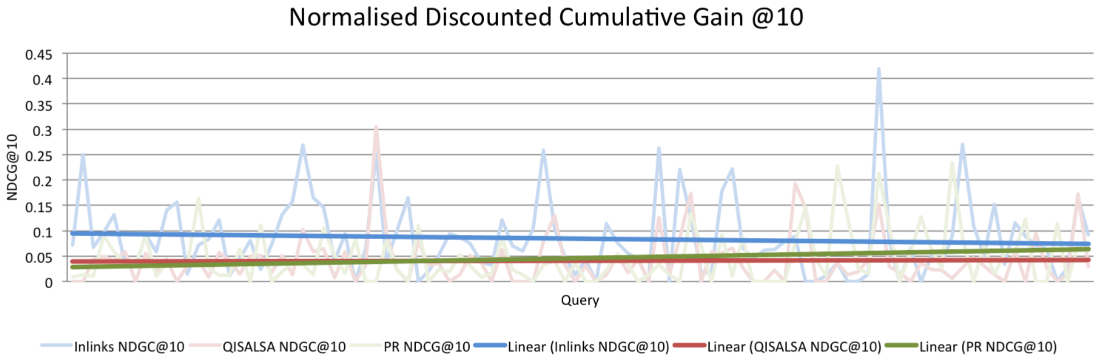


Figure 4.3: Trend of Normalized discounted cumulative gain @10 for each query

The average precision scores for each query for each algorithm are shown for each query in figure 4.4. The variance of these is shown in figure 4.5 which plots a linear trendline for each algorithm accross the resultset. Once again, as per previous results

[18], on average inlinks performs better than PageRank. We find that QISALSA slightly outperforms PageRank for precision on average.

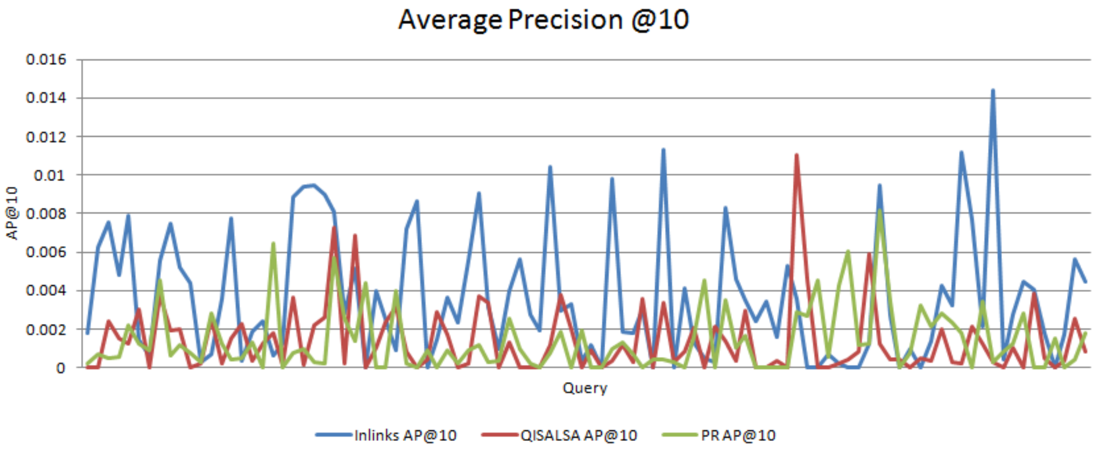


Figure 4.4: Average precision @10 for each query

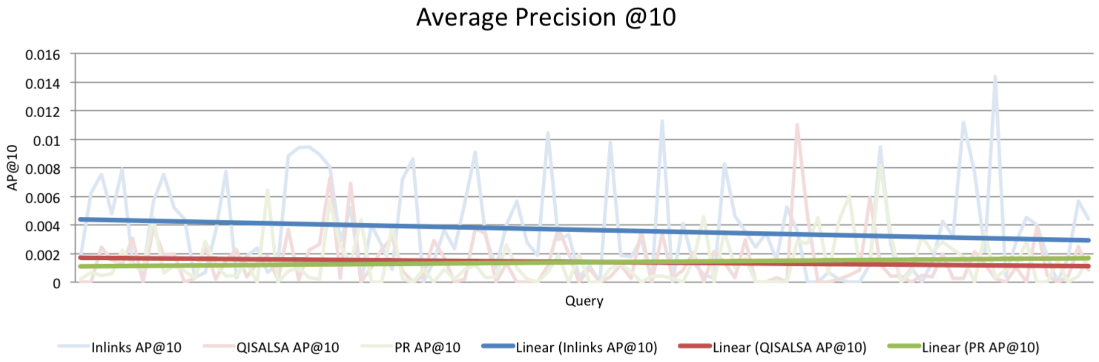


Figure 4.5: Trend of Average precision @10 for each query

The reciprocal rank scores for each query for each algorithm are shown for each query in figure 4.6. The variance of these is shown in figure 4.7 which plots a linear trendline for each algorithm accross the resultset. Once again, as per previous results [18], on average inlinks performs better than PageRank. We find that QISALSA outperforms PageRank for reciprocal rank.

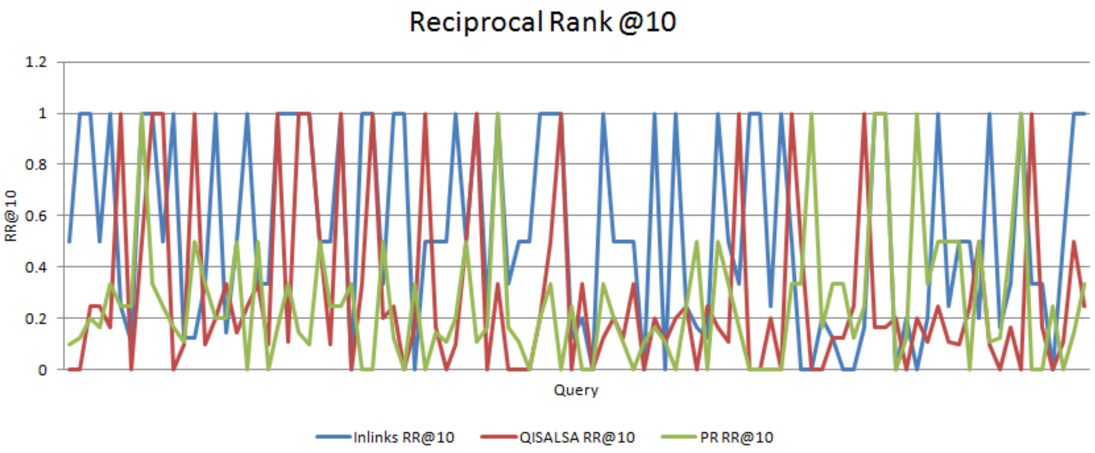


Figure 4.6: Reciprocal rank @10 for each query

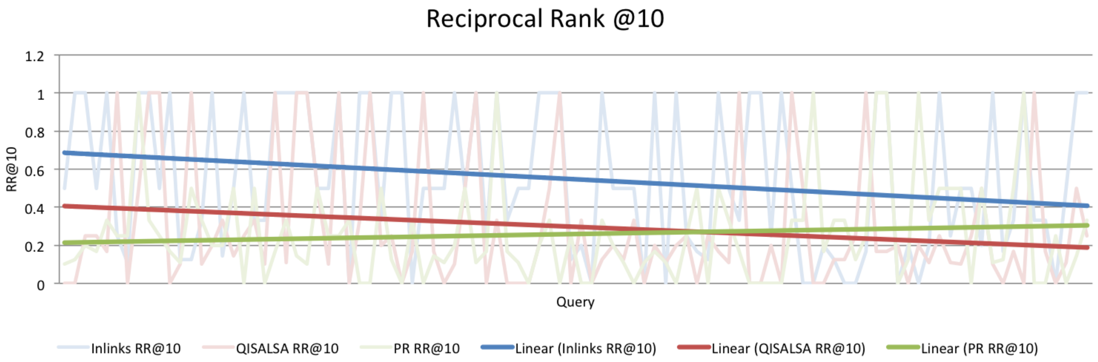


Figure 4.7: Trend of Reciprocal rank @10 for each query

4.7 Discussion

This paper has proposed an algorithm that attempts to combine ideas from PageRank and SALSA. The results suggest that the proposed method of ranking pages does not outperform PageRank by a significant margin and that without improvement may not be appropriate for use in a large-scale search engine. The MRR@10 result with QISALSA outperforming PageRank suggests that QISALSA may provide a better ordering of results based on relevancy. However the NDCG@10 result suggested that despite this beter ordering QISALSA may not produce a results set which is as relevant on average as PageRank. The MAP@10 results show little variance between the two algorithms. We feel that QISALSA requires fundamental improvements to be more effective against comparable ranking methods.

There are a number of future directions to be taken from these results. The first is to combine our formulation of QISALSA with a text-based ranking algorithm such as BM25F [12] to test performance in a more realistic scenario. This could assist with improving the relevancy of the results set which combined with the improved ordering could demonstrate a strong case for the algorithms use. Real world systems commonly create a meta-algorithm in this manner to rank pages. The other possible directions are based around the two more general issues we have addressed. How can we improve the computational (time) efficiency of SALSA-like algorithms and how can we improve the ranking effectiveness of PageRank-like algorithms? As suggested by our results, the best approach may not be to attempt a combination of features from each of these and instead focus on each problem in isolation.

5 Conclusions

We believe that the problem of ranking Internet pages is a fundamental problem that is intriguing and contains a varied array of open problems. These ranking systems are applied in the technology we use every day and are of increasing importance as we create and catalogue more information. Assessing and improving their effectiveness using theoretical and empirical approaches is an open and important area of exploration.

In this thesis we provided a small set of amendments to the work of Altman et al before producing an ordinal, graph-theoretic representation of Edge Weighted PageRank in which the derivation is polynomially bound by the size of the input graph. This contribution provides greater justification for use of PageRank in a weighted environment and lays the groundwork for a rigorous theoretical comparison and evalation of the major search ranking algorithms. It adds to a small but important body of work in the domain of theoretical assessment of ranking algorithms.

We created a new algorithm which combines ideas from the SALSA (Stochastic Approach to Link-Structure Analysis) algorithm with the computational benefits of precalculating a ranking of pages in a query-independent manner. We compared this to other popular approaches: Inlinks, PageRank and HITS. When applied to a small set of web pages this new algorithm didn't outperform the traditional approaches by a significant margin. This result supports the hypothesis that improvements in search ranking algorithms are more easily discovered by combining other methods with link structure analysis. The near future is focused on improvements in textual content analysis, more intelligent query analysis and the combination of other page, domain and user data with results from traditional link structure analysis in meta-algorithms.

Bibliography

[1] Jeffrey S. Rosenthal Panayiotis Tsaparas Allan Borodin, Gareth O. Roberts. Link analysis ranking: Algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5:231 – 297, 2005.

[2] Alon Altman and Moshe Tennenholtz. Ranking systems: The pagerank axioms. Proceedings of the 6th ACM conference on Electronic commerce (EC-05), 2004.

[3] Michael I. Jordan Andrew Y. Ng, Alice X. Zheng. Stable algorithms for link analysis. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001.

[4] Kenneth J. Arrow. *Social Choice and Individual Values (Second Edition)*. Yale University Press, 1963.

[5] Will Hill Brian Amento, Loren Terveen. Does "authority" mean quality? predicting expert quality ratings of web documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000.

[6] Ian Soboro Charles L. A. Clarke, Nick Craswell. Overview of the trec 2009 web track. The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings, 2009.

[7] Ian Soboro Ellen M. Voorhees Charles L. A. Clarke, Nick Craswell. Overview of the trec 2011 web track. In *The Twentieth Text REtrieval Conference (TREC 2011) Proceedings*, 2011.

[8] Ian Soboro Gordon V. Cormack Charles L. A. Clarke, Nick Craswell. Overview of the trec 2010 web track. In *The Nineteenth Text REtrieval Conference (TREC 2010) Proceedings*, 2010.

[9] Huan Chang David Cohn. Learning to probabilistically identify authoritative documents. Proceedings of the 17th International Conference on Machine Learning, 2000.

[10] Massimo Franceschet. Pagerank: standing on the shoulders of giants. *Communications of the ACM*, Volume 54 Issue 6:92–101, 2011.

[11] Taher Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 2003.

[12] Michael Taylor Suchi Saria Stephen Robertson Hugo Zaragoza, Nick Craswell. Microsoft cambridge at trec-13: Web and hard tracks. In *The Thirteenth Text REtrieval Conference (TREC 2004) Proceedings*, 2004.

[13] Haveliwala T.H. Manning C.D. & Golub G.H. Kamvar, S.D. Exploiting the block structure of the web for computing pagerank. Technical report, Stanford University, 2003.

[14] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, 1999.

[15] Any N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton University Press, 2006.

[16] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks*, 33, 2000.

[17] Greg Linden. Marissa mayer at web 2.0, 2006.

[18] Michael Taylor Marc Najork, Hugo Zaragoza. Hits on the web: How does it compare? In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.

[19] Rina Panigrahy Marc Najork, Sreenivas Gollapudi. Less is more: Sampling the neighborhood graph makes salsa better and faster. In *2nd ACM International Conference on Web Search and Data Mining (WSDM)*, 2009.

[20] Zdravko Markov and Daniel T. Larose. *Data Mining The Web.* John Wiley & Sons, 2007.

[21] MA Najork. Comparing the effectiveness of hits and salsa. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007.

[22] Marc Najork and Nick Craswell. Efficient and effective link analysis with precomputed salsa maps. In *17th ACM Conference on Information and Knowledge Management (CIKM)*, 2008.

[23] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[24] Sankar K. Pal. Web mining in soft computing framework: Relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, 13, 2002.

[25] I. Palacois-Huerta and O. Volij. The measurment of intelectual influence. *Econometrica*, 72(3):963–977, 2004.

[26] Richard Zeckhauser Paul Resnick. Trust among strangers in internet transactions: Empirical analysis of ebay' s reputation system. *Volume Advances in Applied Microeconomics*, 11:127 – 157, 2002.

[27] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large altavista query log, 1998.

[28] Marc Najork Sreenivas Gollapudi and Rina Panigrahy. Using bloom filters to speed up hits-like ranking algorithms. In *5th Workshop on Algorithms and Models for the Web Graph (WAW)*, 2007.

[29] Wensi Xi Zheng Chen Yi Liu Michael R. Lyu Wei-Ying Ma Yizhou Lu, Benyu Zhang. The powerrank web link analysis algorithm. Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, 2004.

[30] Ji-Rong Wen Wei-Ying Ma Zaiqing Nie, Yuanzhi Zhang. Object-level ranking: bringing order to web objects. WWW 2005 Proceedings of the 14th international conference on World Wide Web, 2005.