Evaluating Eigenvector Methods for Link Analysis Ranking Under Perturbations

CSci 8363: Final Project Presentation

Aditya Pakki

Department of Computer Science & Engineering
University of Minnesota

December 6, 2017

HITS Algorithm

- Hubs are out-degrees, authorities are in-degrees
- $a^{t+1} = (A^T A)a^t$
  $h^{t+1} = (AA^T)h^t$
- One hyper parameter h^0, trivial choice $[1, 1, \ldots, 1]^T$
- Large eigen gap (δ) + minor perturbation retains ranking
- Converse holds - $\mathcal{O}(\delta)$ perturbation $\implies$ $\Omega(1)$ eigen vector change

PageRank Algorithm

- Probability transition matrix M - normalized rows, $U_{ij} = \frac{1}{n}$
- PageRank(PR) scores $p = (\epsilon U + (1 - \epsilon)M)^T p$
- ϵ typically 0.1-0.2
- If pertrubed pages didn't have high PR scores, $p \approx \widetilde{p}$

Perturbation Results

- CORA dataset ranking for AI papers, for 5 runs
- HITS on the left; PageRank on the right for $\epsilon = 0.2$

1	"Genetic algorithms in search, optimization...", Goldberg	1	3	1	1	1
2	"Adaptation in natural and artificial systems", Holland	2	5	3	3	2
3	"Genetic programming: On the programming of...", Koza	3	12	6	6	3
4	"Analysis of the behavior of a class of genetic...", De Jong	4	52	20	23	4
5	"Uniform crossover in genetic algorithms", Syswerda	5	171	119	99	5
6	"Artificial intelligence through simulated...", Fogel	6	135	56	40	8
7	"A survey of evolution strategies", Back+al	10	179	159	100	7
8	"Optimization of control parameters for genetic...", Grefenstette	8	316	141	170	6
9	"The GENITOR algorithm and selection pressure", Whitley	9	257	107	72	9
10	"Genetic algorithms + Data Structures = ...", Michalewicz	13	170	80	69	18
11	"Genetic programming II: Automatic discovey...", Koza	7	-	-	-	10
2060	"Learning internal representations by error...", Rumelhart+al	-	1	2	2	-
2061	"Learning to predict by the method of temporal...", Sutton	-	9	4	5	-
2063	"Some studies in machine learning using checkers", Samuel	-	-	10	10	-
2065	"Neuronlike elements that can solve difficult...", Barto+Sutton	-	-	8	-	-
2066	"Practical issues in TD learning", Tesauro	-	-	9	9	-
2071	"Pattern classification and scene analysis", Duda+Hart	-	4	7	7	-
2075	"Classification and regression trees", Breiman+al	-	2	5	4	-
2117	"UCI repository of machine learning databases", Murphy+Aha	-	7	-	8	-
2174	"Irrelevant features and the subset selection...", John+al	-	8	-	-	-
2184	"The CN2 induction algorithm", Clark+Niblett	-	6	-	-	-
2222	"Probabilistic reasoning in intelligent systems", Pearl	-	10	-	-	-

1	"Genetic Algorithms in Search, Optimization and...", Goldberg	1	1	1	1	1
2	"Learning internal representations by error...", Rumelhart+al	2	2	2	2	2
3	"Adaptation in Natural and Artificial Systems", Holland	3	5	6	4	5
4	"Classification and Regression Trees", Breiman+al	4	3	5	5	4
5	"Probabilistic Reasoning in Intelligent Systems", Pearl	5	6	3	6	3
6	"Genetic Programming: On the Programming of ...", Koza	6	4	4	3	6
7	"Learning to Predict by the Methods of Temporal ...", Sutton	7	7	7	7	7
8	"Pattern classification and scene analysis", Duda+Hart	8	8	8	8	9
9	"Maximum likelihood from incomplete data via...", Dempster+al	10	9	9	11	8
10	"UCI repository of machine learning databases", Murphy+Aha	9	11	10	9	10
11	"Parallel Distributed Processing", Rumelhart+McClelland	-	-	-	10	-
12	"Introduction to the Theory of Neural Computation", Hertz+al	-	10	-	-	-

Stochastic Approach for Link Structure Analysis(SALSA)

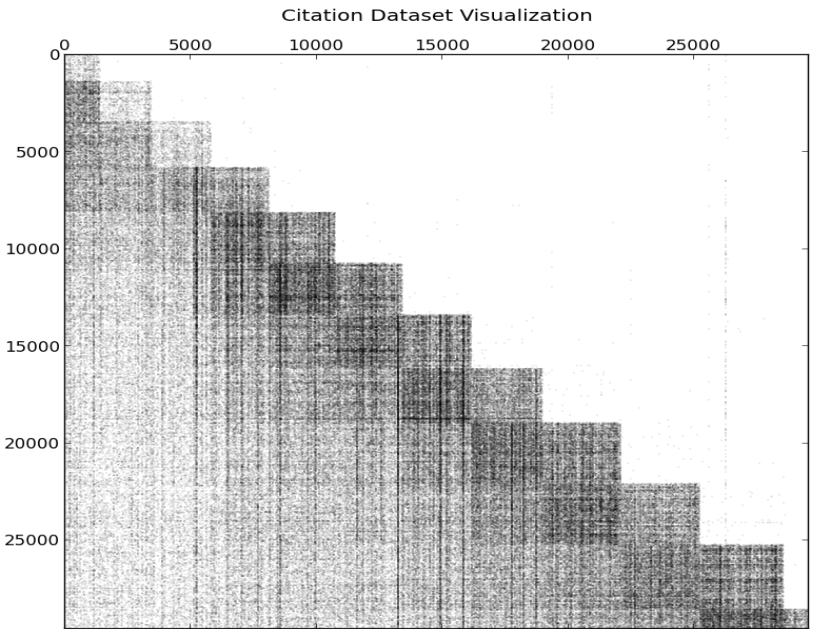
- Random Walk from PR, hubs V_h & authorities V_a from HITS
- Two random Walks; forward: $V_a \rightarrow V_h$, backward: $V_h \rightarrow V_a$
- $a^{t+1} = A_c^T A_r a^t$
  $h^{t+1} = A_r A_c^T h^t$
- Two hyper parameters - a^t depends on a^0, h^t depends on h^0

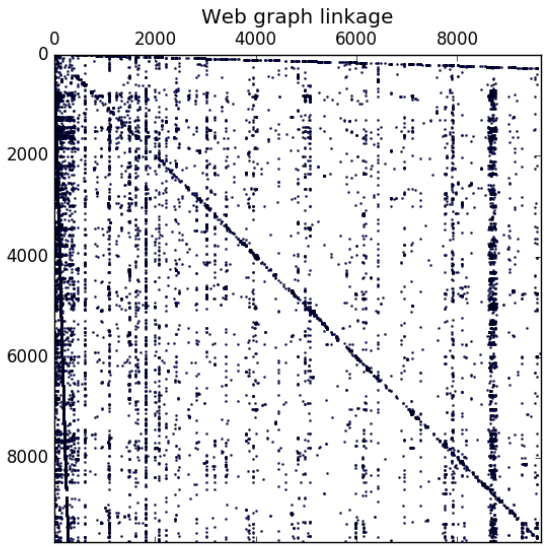
Randomized HITS

- Jumps to random page with ϵ; forward, backward random walks with $(1 - \epsilon)$
- $a^{t+1} = \epsilon\vec{1} + (1 - \epsilon)A_r^T h^t$
  $h^{t+1} = \epsilon\vec{1} + (1 - \epsilon)A_c a^{t+1}$
- Two hyper parameters - ϵ and h^0

Experiment Datasets

- Rank datasets for citation & web graphs under perturbation
- High Energy Physics Citation dataset [1]
  29554 nodes, **352807** edges
- Web graph hyperlink dataset [2]
  9663 nodes, **16151** edges

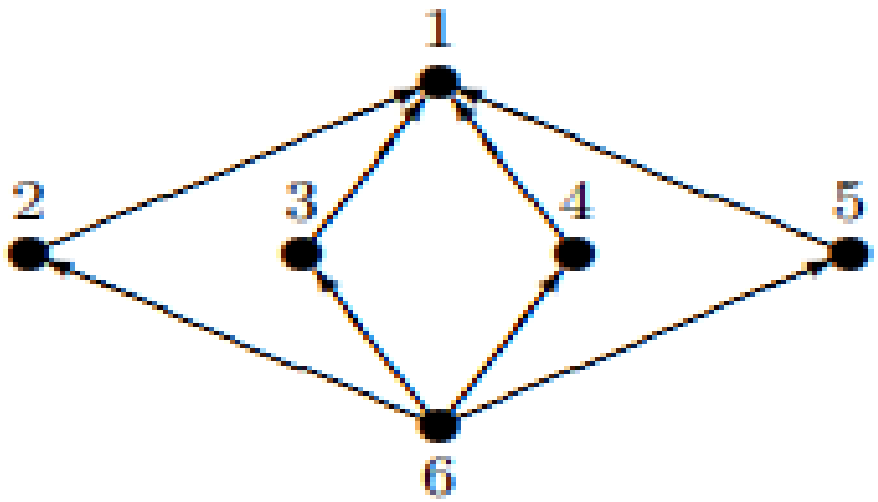




[1] https://snap.stanford.edu/data/cit-HepPh.html
[2] http://www.cs.cornell.edu/courses/cs685/2002fa/data/gr0.California

Impact of Initial Vector

- Ranking depends on the principal eigen vector, smaller δ flips Eigen vectors v_1, v_2
- For simplicity, pick uniform values $[1, 1, \ldots, 1]^T, [\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}]^T, [\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}]^T$
- HITS & SALSA gives inconsistent hubs, authority scores for uniform initial vector - Good hubs point to good authorities
- Account for the weights, for below graph in [1], pick $h = [0, 0.125, 0.125, 0.125, 0.125, 0.5]^T$



Tweaking parameters

- Evaluate PageRank, RandomHITS for ϵ range 0.1 - 0.5
- Deleting random datasets leads to disconnected components, run BFS to identify connectivity
- Perturbation Ranking on 10, 20, 30, 40% deleted datasets

Web Hyper Links Dataset

- Average ranking of 4 algorithms on 30% perturbed dataset, favorable to HITS
- $\epsilon = 0.15$; HITS, PageRank - python networkx modules

Orig	HITS	PR	RHITS	SALSA	Webpages
1	1	13	4	29	http://www.ca.gov/
2	3	28	7	460	http://www.sen.ca.gov/
3	9	15	15	216	http://www.assembly.ca.gov/
4	2	19	3	261	http://www.leginfo.ca.gov/calaw.html
5	91	2883	2883	2883	http://www.yahoo.com/
6	4	61	19	614	http://www.house.gov/
7	12	8864	8864	8864	http://www.fedworld.gov/
8	5	122	50	273	http://www.lao.ca.gov/
9	10	73	44	322	http://www.dot.ca.gov/
10	6	286	143	146	http://www.courtinfo.ca.gov/
11	7	46	10	820	http://www.epa.gov/
12	8	126	41	1338	http://www.census.gov/
13	13	1554	1554	1554	http://www.berkeley.edu/
14	50	2884	2884	2884	http://www.lycos.com/
15	11	56	83	231	http://www.ss.ca.gov/
16	48	2	46	925	http://www.caltech.edu/
17	14	50	16	831	http://goldmine.cde.ca.gov/
18	27	51	30	399	http://www.excite.com/
19	17	247	128	263	http://www.ftb.ca.gov/
20	51	32	165	993	http://www.csun.edu/

Web Hyper Links Dataset

- Average ranking of 5 runs, 4 algorithms on 25% perturbed citation dataset
- $\epsilon = 0.15$; PageRank works accurately for most of the deletions

Orig	HITS	PR	RHITS	SALSA	Papers
1	244	1	69	22847	Noncompact Symmetries in String Theory
2	11	2	2	76	Monopole Condensation, And Confinement In N=2 Supersymmetric Yang-Mills
3	12131	3	2175	14309	An Algorithm to Generate Classical Solutions for String Effective Action
4	7	4	10	4989	String Theory Dynamics In Various Dimensions
5	5	5	9	90	Dirichlet-Branes and Ramond-Ramond Charges
6	379	6	35	2330	Exact Results on the Space of Vacua of Four Dimensional SUSY Gauge
7	1	7	1	16619	The Large N Limit of Superconformal Field Theories and Supergravity
8	10	8	17	3569	Unity of Superstring Dualities
9	17	9	8	22983	Monopoles, Duality and Chiral Symmetry Breaking in N=2 Supersymmetric
10	118	10	46	878	Strong-Weak Coupling Duality in Four Dimensional String Theory
11	1174	11	25	20897	ERRATA for "One-Loop Threshold Effects in String Unification"
12	6	12	7	241	M Theory As A Matrix Model: A Conjecture
13	445	13	61	19241	Supersymmetry as a Cosmic Censor
14	9	14	15	238	Bound States Of Strings And p-Branes
15	299	15	58	22189	Duality Symmetric Actions
16	2	16	3	1456	Anti De Sitter Space And Holography
17	54	17	20	22381	Target Space Duality in String Theory
18	671	18	433	4723	Electric Magnetic Duality in String Theory
19	3	19	4	22650	Gauge Theory Correlators from Non-Critical String Theory
20	59	20	66	349	Massless Black Holes and Conifolds in String Theory

In Essence

- Evaluated various EigenVector methods for Ranking Hyperlink Web graph, citation datasets
- Randomized HITS, PageRank are stable under perturbations for Citation dataset
- To Do - Study the impact of Initial Vector on HITS, SALSA
- To Do - Compare results with theorems from paper [4]

References I

Farahat, A., LoFaro, T., Miller, J. C., Rae, G., and Ward, L. A.
Authority rankings from hits, pagerank, and salsa: Existence, uniqueness, and effect of initialization.
SIAM Journal on Scientific Computing 27, 4 (2006), 1181–1201.

Kleinberg, J. M.
Authoritative sources in a hyperlinked environment.
J. ACM 46, 5, 604–632.

Lempel, R., and Moran, S.
The stochastic approach for link-structure analysis (salsa) and the tkc effect1abridged version1.
Computer Networks 33, 1 (2000), 387 – 401.

References II


Ng, A. Y., Zheng, A. X., and Jordan, M. I.
Link analysis, eigenvectors and stability.
In *International Joint Conference on Artificial Intelligence*, vol. 17, LAWRENCE ERLBAUM ASSOCIATES LTD (2001), 903–910.


Page, L., Brin, S., Motwani, R., and Winograd, T.
The pagerank citation ranking: Bringing order to the web.
Technical Report 1999-66, Stanford InfoLab, November 1999.
Previous number = SIDL-WP-1999-0120.