Score and Rank Convergence of HITS

Enoch Peserico and Luca Pretto
Dip. Ing. Informazione, Univ. Padova, Italy
{enoch,pretto}@dei.unipd.it

ABSTRACT

How many iterations does the (ever more) popular HITS algorithm require to converge in score and, perhaps more importantly, in rank (i.e. to get the nodes of a graph "in the right order")? After pinning down the elusive notion of convergence in rank we provide the first non-trivial bounds on the convergence of HITS. A "worst case" example, requiring a number of iterations superexponential in the size of the target graph to achieve even "mild" convergence, suggests the need for greater caution in the experimental evaluation of the algorithm - as recent results of poor performance (e.g. vs. SALSA) might be due to insufficient iterations, rather than to an intrinsic deficiency of HITS. An almost matching upper bound shows that, as long as one employs exponential acceleration e.g. through a "squaring trick", a polynomial running time (practical in many application domains) always provides strong convergence guarantees.

Categories and Subject Descriptors: F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems - Computations on discrete structures

General Terms: Algorithms, Theory

Keywords: link analysis, HITS, convergence, rank

1. HITS

Kleinberg's celebrated HITS algorithm [4] is probably the most widely used link analysis algorithm outside of Web search - making it a reference algorithm for today's link analysis, much like quicksort for sorting. Originally proposed to rank Web pages, HITS has been subsequently employed, sometimes with small variations, to rank graph nodes in a vast (and growing!) number of IR applications, often with little or no connection to Web search (e.g. [1, 6, 7, 3] - see [10] for a more comprehensive bibliography).

The original version of HITS works as follows. In response to a query, a search engine first retrieves a set of Web graph nodes on the basis of pure textual analysis; for each such node it also retrieves all nodes pointed by it, and up to d nodes pointing to it. HITS operates on the subgraph G induced by this *base set* (which clearly depends on the query) iteratively (re)computing an *authority score vector* $\mathbf{a}^{(t)}$:

$$\mathbf{a}^{(0)} = \frac{\mathbf{A}^T\mathbf{1}}{\|\mathbf{A}^T\mathbf{1}\|_2} \qquad \mathbf{a}^{(t)} = \frac{(\mathbf{A}^T\mathbf{A})\mathbf{a}^{(t-1)}}{\|(\mathbf{A}^T\mathbf{A})\mathbf{a}^{(t-1)}\|_2} \qquad t = 1, 2, \ldots$$

where $\mathbf{A}$ is the adjacency matrix of G and $\mathbf{1}$ denotes the (column) vector whose components are all 1. The i^{th} component a_i of the limit score vector $lim_{t\to\infty}\mathbf{a}^{(t)}$ (which always exists [4]) is the *authority* score of the i^{th} node, that summarizes both its quality and its relevance to the query.

2. SCORE AND RANK CONVERGENCE

HITS effectively computes the dominant eigenvector of $\mathbf{A}^T\mathbf{A}$ using the Power Method [2]; thus its speed of convergence in score is tied to the separation of the first and second eigenvalues of that matrix. However, no bounds on this separation are known *for the matrix derived from a graph* - for arbitrary matrices of any fixed size the separation can be arbitrarily small and convergence arbitrarily slow.

In fact, most applications employ the score vector directly to *rank* the nodes of the target graph. In these cases it is even more important to understand the number of iterations required by HITS to converge *in rank* - informally to assign scores that, although potentially quite different from the limit scores, still place nodes in a "correct" order. Again, no bounds are known on the convergence of HITS in rank: only a few experimental results are available, and only for the Web graph [4]. Being heavily application-dependent, these provide little information to guide the researcher who would port the famous algorithm to new application domains.

The informal definition of rank convergence above has two flaws. First, it does not deal with ties or "almost ties": if the difference between the limit scores of two nodes is negligible, an algorithm effectively converges to a "correct" ranking even if it keeps switching their relative ranks. Second, it does not distinguish between an algorithm failing for a long time to produce a ranking even remotely close to the ultimate ranking, and one taking a long time to reach the ultimate ranking, but quickly reaching a ranking "close" to it (e.g. with all elements correctly ranked, save the last few).

To address these issues we introduce a more general and formal definition of convergence in rank. Let a ranking be *compatible* with a score vector if no node with a higher score is ranked worse than one with a lower score (ties can be broken arbitrarily). Then:

DEFINITION 1. *Let an iterative algorithm ALG produce at each iteration t a score vector $\mathbf{v}(t)$ and converge to a score vector $\mathbf{v}(\infty)$. Then ALG ϵ-converges on h of the top k ranks in (at most) τ steps if, for all iterations $t \geq \tau$, at least h of the top k items in a ranking compatible with $\mathbf{v}(t)$ are also among the top k items in a ranking compatible with $\mathbf{v}(\infty)$, or compatible with some vector $\mathbf{w}(t)$ at distance at most ϵ from $\mathbf{v}(\infty)$.*

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

In other words, we assume an algorithm has converged in rank as soon as it "gets right" (and keeps getting right) at least h of the top k items of any ranking compatible with the limit score vector, or with any score vector "sufficiently close" to it (the definition above implicitly assumes some distance function between score vectors - e.g. $\|\cdot\|_2$). Note that with $\epsilon = 0$ and $h = k\ \forall k$ our definition collapses back to the stricter, "naive" definition of convergence in rank. Our tech report [10] provides a more thorough discussion of the "nice" properties of this definition in the context of existing literature.

3. SLOWLY, BUT NOT TOO SLOWLY

In our tech report [10] we prove that HITS can converge slowly, but not too slowly, both in score and in rank. On the one hand, we prove that a number of iterations superexponential in the size of the target graph might be necessary to converge to a ranking (and a score) that is even remotely accurate (see Theorem 1 below). On the other hand, we prove that such a number of iterations is also sufficient (Theorems 2 and 3). In this regard, one should note that the Power Method can be exponentially accelerated with only a modest space overhead through *repeated squaring* [5], computing $\mathbf{a}^{(t)} = (\mathbf{A}^T\mathbf{A})^t\mathbf{A}^T\mathbf{1}$ by first computing $(\mathbf{A}^T\mathbf{A})$, $(\mathbf{A}^T\mathbf{A})^2$, $(\mathbf{A}^T\mathbf{A})^4$, $\ldots$, $(\mathbf{A}^T\mathbf{A})^{2^{\lfloor \lg(t)\rfloor}}$ and then multiplying an appropriate subset of those $\lfloor \lg(t)\rfloor$ matrices. Thus, the complexity of converging to a result extremely close to the limit ranking (and score) is at most polynomial in the number n of nodes of the target graph: more precisely $O(n^{4+\mu})$ (where $\Theta(n^{2+\mu})$ is the complexity of n by n matrix multiplication) and $O(n^{3+\mu})$ in the important case of authority connected graphs. These results can be easily generalized to weighted graphs (important for many applications of HITS, e.g. [7, 6]) with only mild restrictions on the link weights. [10] provides a complete discussion of the hypotheses and proof techniques involved in the three Theorems below.

THEOREM 1. *For all $k \geq 3$ and $s \geq 3$ there exists a(n authority connected) graph Γ of maximum degree $4k$ and $3(k+1)s + k^3 + 2k^2 + 2k + 2 \approx k^3 + 3ks$ nodes on which HITS fails to ϵ-converge on more than $k+1$ of the top k^2+k ranks in less than $k^{\Theta(s)}$ iterations for all $\epsilon \leq \bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$.*

THEOREM 2. *Let G be a graph of n nodes and maximum degree g whose arcs have* integer *weights at least 1 and at most w. Denote by $\mathbf{A}$ the weighted adjacency matrix of G. If $\mathbf{A}^T\mathbf{A}$ is a block matrix with at least two blocks of size m_1 and m_2 whose dominant, positive eigenvalues λ_1 and $\lambda_2 < \lambda_1$ are respectively the largest and second largest eigenvalue of $\mathbf{A}^T\mathbf{A}$, then HITS ϵ-converges in score (in $\|\cdot\|_2$), and therefore on all ranks, on the nodes of G in at most $(wg)^{O(m_1 m_2)}(\lg(\frac{1}{\epsilon}) + \lg(n))$ iterations.*

THEOREM 3. *Let G be a graph of n nodes and maximum degree g whose arcs have weights at least 1 and at most w. Denoting by $\mathbf{A}$ the weighted adjacency matrix of G, if $\mathbf{A}^T\mathbf{A}$ is a block matrix such that all its non-zero blocks have size at most m and if the largest and the second largest eigenvalues of $\mathbf{A}^T\mathbf{A}$ are relative to the same block (including the case of just one non-zero block, i.e. if G is authority connected), then HITS ϵ-converges in score (in $\|\cdot\|_2$), and therefore on all ranks, on the nodes of G in at most $(wg)^{O(m)}(\lg(\frac{1}{\epsilon}) + \lg(n))$ iterations.*

4. CONCLUSIONS

The vast and growing number of applications of HITS in IR suggests greater effort should be spent in studying its convergence in score and, perhaps more importantly, *in rank* (the latter being a far more slippery issue - indeed, our contribution of a precise definition of the problem should be valuable in the study of other iterative ranking algorithms). We provide the first non-trivial bounds on the convergence rate of HITS, both in score and in rank.

We provide a "worst case" example in which HITS requires a number of iterations superexponential in the graph size to converge, not only to within a relatively large distance of the limit score vector, but also to more than a tiny fraction of the limit top ranks. This suggests that experimental evaluation of the algorithm should be conducted with great(er?) care. Approximately half of all HITS related papers in the top IR conferences fail to report the termination criteria with which the algorithm is run, and most remaining ones report stopping the computation when the difference between two consecutive score vectors falls below a certain threshold - a condition that by itself in no way guarantees a correspondingly small distance from the limit score and/or rank vector. *One should then make certain that reported instances of poor performance of HITS (e.g. [8, 9]) are indeed due to an intrinsic deficiency of the algorithm, rather than simply to an insufficient number of iterations*; although understanding the properties of HITS when stopped before convergence (sometimes the only feasible option, as this work shows) certainly seems a very promising line of research.

We also provide an almost matching upper bound on the convergence rate of HITS. While rather weak, it does show that, *as long as HITS is accelerated through e.g. repeated squaring* (a well known scheme that preserves almost all the "nice" properties of the Power Method) *a polynomial running time guarantees almost complete convergence.* In practical terms, this means HITS might be run online with strong convergence guarantees on graphs of up to several hundred nodes, and up to several thousands or even tens of thousands in the case of offline computation and/or "favourable" graph topologies; remembering that base sets are typically quite small, this includes many applications of practical interest.

5. ACKNOWLEDGEMENTS

This work was supported by EU IP AEOLUS (FP6-015964).

6. REFERENCES

[1] M. W. Berry, ed. *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer, NY, 2004.

[2] G. H. Golub and C. F. Van Loan. *Matrix Computations, 3rd ed.,* J. Hopkins Un. Press, 1996.

[3] P. Jurczyk and E. Agichtein. HITS on question answer portals: Exploration of link analysis for author ranking. *SIGIR'07.*

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5), 1999.

[5] D. E. Knuth. *The Art of Computer Programming*, vol. 2. 3rd ed., Addison-Wesley, 1998.

[6] O. Kurland and L. Lee. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. *SIGIR'06.*

[7] S. Mizzaro and S. Robertson. HITS hits TREC - exploring IR evaluation results with network analysis. *SIGIR'07.*

[8] M. Najork. Comparing the effectiveness of HITS and SALSA. *CIKM'07.*

[9] M. Najork, H. Zaragoza, and M. Taylor. HITS on the Web: How does it compare? *SIGIR'07.*

[10] E. Peserico and L. Pretto. *HITS can converge slowly, but not too slowly, in score and rank.* Un. Padova Tech. Rep., 2009.