Analyzing Ranking Algorithms

Izabela Karennina Travizani Maffra

Universidade Federal de Minas Gerais
Computer Science Department
Complex Networks

1 Introduction

When it comes to information retrieval systems, the ranking component plays a crucial role. Even though the necessity of ranking results is an old issue, after the explosion of the Web it gained a very particular importance. A Web user wants to be able to search the huge amount of information that is available in the Web and still find easily and quickly the most relevant pages to the topic. If a search engine is not capable of showing the most important results in the first page or so, it is deemed to be useless.

In the field of Link Analysis, many ranking algorithms have been proposed. Among the most famous ones, one can cite:

- HITS (Hyperlink-Induced Topic Search) [2], which was proposed in 1999 by Jon Kleinberg. In this algorithm, each page is assigned two scores, known as *hub* and *authority*. The definition of those terms are mutually recursive: in general lines, a good hub is a page that points to many authority pages, whereas a good authority is a page which is pointed by many good hubs.

  These scheme stems from the original organization of the Web, in which there were many pages that resembled a catalog and acted as real hubs: they simply contained lists of links to other web pages.

- PageRank [4], which was developed by Brin and Page in 1996. The basic idea behind PageRank is that web pages can be ordered in a hierarchy by "link popularity": a page is ranked higher as there are more links to it. Shortly after publishing PageRank, Page and Brin founded Google Inc., the company behind the Google search engine.

Many adaptions and improvements have been proposed to these original algorithms. In the end, the actual ranking decisions made by popular search engines such as Google are highly secret and reason for a lot of speculation. After all, being on top of the web searches may very often result in a boost in profit for many businesses.

In this work we are going to compute and evaluate the results for both HITS and PageRank, when applied to the Stanford web graph, as of 2002, in which nodes represent pages from Stanford University (`stanford.edu`) and directed edges represent hyperlinks between them. These dataset is available in `http://snap.stanford.edu/data/web-Stanford.html`.

2 Results

We computed PageRank and HITS scores using the `networkx` library[1].

The Stanford web graph contains 281903 nodes, which means that a manual inspection of the results is unfeasible. In order to analyze the results, we first compare the top 20 nodes for each of the algorithms. These comparison can be seen in Tables 1, 2 and 3. For each top result provided by one of the algorithms, we provide the position of that node in the other rankings.

	Position in rank	
PageRank	HITS: hub	HITS: authority
1	100678	24123
2	47001	**1**
3	115188	36953
4	110929	43570
5	107853	29153
6	19317	**2**
7	215222	190374
8	216777	187295
9	18996	**3**
10	135957	37167
11	138884	45446
12	70712	45247
13	70824	16082
14	215225	190373
15	215223	190371
16	215224	190372
17	225738	200215
18	110925	45220
19	138885	43591
20	41707	**7**

Table 1: The corresponding position for the top-20 results according to PageRank, when considering the hub score and the authority score provided by HITS.

[1] `http://networkx.github.io/`

Position in rank	
HITS: authority	PageRank
1	69835
2	87244
3	91574
4	91060
5	12225
6	12712
7	67406
8	68789
9	4262
10	3120
11	90302
12	3449
13	93240
14	160
15	264086
16	268904
17	270041
18	268828
19	205
20	64758

Table 2: The corresponding position for the top-20 hubs according to HITS, when considering PageRank score.

When we analyze these tables, we can see that the results are considerably different. A first conclusion that we might draw is that apparently the most unrelated scores would be the PageRank and authority. In Table 2, one can see that most of the top-20 authorities are left very far behind by PageRank.

When we compare the authority score and PageRank, by analyzing Tables 1 and 2, the results are a little bit more alike, which is reasonable. In the end, these scores represent directly the result that is delivered for the user by HITS and PageRank, respectively.

We also compared the ranks provided by the algorithms using the **Kendall tau rank correlation coefficient** [5], which measures the similarity of different rankings of the same data. Formally, it is defined as follows:

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{\frac{1}{2}n(n-1)}$$

A pair is said to be *concordant* if the ranks for both elements agree, i.e., if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

Position in rank	
HITS: hub	PageRank
1	**2**
2	**6**
3	**9**
4	24
5	32
6	36
7	**20**
8	22
9	59
10	46
11	122
12	213
13	238
14	239
15	27
16	76
17	168
18	176
19	162
20	182

Table 3: The corresponding position for the top-20 hubs according to HITS, when considering PageRank score.

The denominator is the total number of pair combinations, so the coefficient must be in the range $-1 \leq \tau \leq 1$.

This coefficient holds the following properties:

(i) If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.

(ii) If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value -1.

(iii) If X and Y are independent, then we would expect the coefficient to be approximately zero.

In Table 4, we can see the results for the calculation of the coefficient. When we compare the rank generated by PageRank with the rankings generated by both the hub score and authority score in HITS, the τ coefficient is very close to zero, meaning that the rankings are independent. When we consider only the top-200 results for each ranking (which is reasonably much more relevant), the correlation between authority and PageRank increases to 0.078392, but even so this number is still very small.

ranks			τ
PageRank	*vs*	HITS: authority	0.001783
PageRank	*vs*	HITS: hub	-0.002490

Table 4: Kendall tau rank correlation coefficient for the generated rankings.

Without any more details about the pages that each node actually represents, it is difficult to make any further analyses. In a ideal scenario, we would be able to perform queries and compare the ranking of the results.

3 SALSA: Combining HITS and PageRank

Stochastic Approach for Link-Structure Analysis (SALSA) [3] is a web page ranking algorithm designed by R. Lempel and S. Moran to assign high scores to hub and authority web pages based on the quantity of hyperlinks among them.

The SALSA algorithm performs a random walk on the bipartite hubs and authorities graph, alternating between the hub and authority sides. The random walk starts from some authority node selected uniformly at random. The random walk proceeds by alternating between backward and forward steps. When at a node on the authority side of the bipartite graph, the algorithm selects one of the incoming links uniformly at random and moves to a hub node on the hub side. When at a node on the hub side, the algorithm selects one of the outgoing links uniformly at random and moves to an authority. The authority weights are defined to be the stationary distribution f this random walk [1].

Structurally, SALSA resembles HITS, because the same bipartite graph structure of hubs and authorities is considered and those two scores are computed. The difference from HITS relies on how the scores of the authorities are updated in each step. Whereas in HITS the score of the hub is simply broadcast to all the authorities it points to, in SALSA this score is equally divided among the authorities it points to, which is an approach that is closer to PageRank's strategy. Also like PageRank, SALSA's scores are computed by simulating a random walk through a Markov chain that represents the graph of web pages.

In the end, SALSA can be considered an improvement over HITS. It still presents one of the main disadvantages of HITS, which is its query dependency (after each query, a subgraph containing just the nodes more relevant to the topic should be generated). This can impact greatly in the response time of a search engine.

Nevertheless, SALSA behaves better than HITS in the presence of a Tightly Knit Community (TKC), a topological structure within the Web consisting of a small set of pages highly interconnected, which could even represent a purposeful bias to ranking algorithms. In HITS, TKCs tend to affect in a negative way the detection of good authorities.

References

[1] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. Link analysis ranking: Algorithms, theory, and experiments. *ACM Trans. Internet Technol.*, 5(1):231–297, February 2005.

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

[3] R. Lempel and S. Moran. Salsa: The stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, April 2001.

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

[5] Wikipedia. Kendall tau rank correlation coefficient — Wikipedia, the free encyclopedia, 2014. [Online; accessed 4-April-2004].