Ben Levine

11/18/17

Link analysis and web search

**Voting by in-links:** rank pages of any given query by number of sites that link to it

* text search of web is easily gamed, lots of fraud

**Lists:** each website is a list of links (these lists can also be called **hubs**). These lists can be ranked by the sum of the votes received by all pages (also called an **authority**) for which it voted. We can then use this ranking to weight the votes for each page, getting a more refined estimate for what the best page is (we can then use these to re-rank the lists). Iterating through this process is the **principle of repeated improvement**

Each page has separate rankings as a hub and as an authority.

**Authority update rule:** for each page, update the authority score to be the sum of the hub scores that point to it

**Hub update rule:** for each page, update hub to be the sum of the authority scores of all the pages that it points to.

Choose k steps, and sequentially update authorities, then hubs to get a (normalized) set of scores. normalized values will converge to limits, stabilizing over time (DMR to iteration).

**PageRank**

Endow all nodes with 1/n PR, then:

**Basic PageRank update rule:** each page divides its current page rank equally across its outgoing links and passes these equal shares to the pages it points to. Each page updates its new PageRank to be the sum of the shares it receives

* PR in state of entropy; never created or destroyed, just moving around the network

**Scaled PR update rule**: first apply basic PR update rule. then scale down all PR values by a factor (b/t 0,1, in practice between .8 and .9) *s*. divide the residual 1-*s* units of PR equally over all nodes, giving (1-*s*)/n to each

* evaporating water analogy

**Modern Web search**

**Anchor text:** in-line hyperlinked text, frequently used as a match to search terms, keywords

SEO makes web search a moving target as people/firms try to game the system

* search advertising

**Applications beyond the web:**

* Impact factor: average number of citations received by a paper in the given journal over the last two years
* Journal influence weights: applies the principles of PageRank

**Spectral analysis of PageRank:**

Matrix ***N***, ***N***ij specifies the portion of i’s PageRank that should be passed to j in one update step.

Define ***r*** as a vector equal to the PageRank of all nodes:

***r(t)*** = (***N***)’***r(t-1)***

Can do the same with the scaled PageRank update rule, where ***Ñ***ij = *s****N***ij + (1-*s*)/n. Again,

***r(t)*** = (***Ñ***)’***r(t-1)***

Repeated improvement: given that r will converge to limit, at some point: ***r(t)*** = (***Ñ***)’***r(t)***. That is, r will be an eigenvector of ***Ñ***’, with a corresponding eigenvalue of 1.

**Perron’s Theorem:**

Any matrix **P** in which all entrees are positive has the following properties:

* P has a real eigenvalue c>0 such that c>|c’| for all other eigenvalues c’
* there is an eigenvector y with positive real coordinates corresponding to the largest eigenvalue c, and y is unique up to multiplication by a constant
* if the largest eigenvalue c is equal to 1, then, for any starting actor x=/=0 with nonnegative coordinates, the sequence of vectors

Perron’s theorem tells us that there is a unique vector y that remains fixed under the application of the scaled update rule, and that repeated application of the update rule from any starting point converges to y. This vector y thus corresponds to the limiting PageRank values we have been seeking