file/after/tracklang.stypackage/after/bidipackage/after/luabidi

Link Analysis

The problem of quantifying and analyzing online knowledge is important. How do we find trustable web pages? The hyperlinks of a website are helpful.

Something like the **impact factor** for a journal measures the citations per item in a given timeframe. We can have C/N, where C is the numbe rof citations in a given time interval and N is the total number of citations in the previous timeframe.

The number of **inlinks**, or hyperlinks pointing to a site, tell us about how important a site can be. Some of the early algorithms took a look at how many other authoritative sites point to a particular site and how many other sites it points to.

In the **HITS** algorithm (Hypertext Induced Topic Selection) classifies the **hubness** $h(v)$ of a site v and the **authority** $a(v)$ of a site.

When performing a query on a set of sites, we would like to have a mix of relevant (?) and authoritative sites. In this algorithm, we rank the web pages by their in-degree (incoming nodes) and pick the top k to serve as our **root set**. From the root set, we then expand our query to the base set, We then expand the root set into the base set by moving through the children and parents of the root set and adding them to our graph to become the base set.

We repeat the calculation for authoritativeness and hubness of the links until their sum is less than some value ϵ.

One of the most obvious problems with this approach to link classificaiton is that sites may link to other sites to purosefully game this algo rather than because teh linked site is good. If there are enough evil actors that act in this way there could be a network of nodes that appears hub-like and authoritative without being so. Also, sites contain user content, and thus there are many links which might not point to sites that would otherwise be linked. As the iterations go on, the number of authorities and hubs usually converges. Also, some pages contain links to the same other website in multiple places (i.e. link to social media profile in header and footer of the page?).

PageRank

In this algo, we use the rank of the parents and the amount of children to calculate the rank.

$$r(v) = \alpha \sum_{w \in pa[v]} \frac{r(w)}{|ch[w]|} \tag{1}$$

This means the rank of any one site depends on how many children it points to as well. Thus the page would do best to have a few and more important links rather than a lot of unimportant ones.

For a graph, we could have the matrix of outlinks and inlinks for each page, maybe a matrix with a 1 for each link that points to the page in the inlinks and a 1 for each outgoing link in the page for the outlinks. The final output of the PageRank algorithm is the probability that a person taking a random walk through the graph by clicking random links will end up at a specific page, called the **damping factor**. A damping factor of 0.85 would mean a surfer has a 0.85 chance of continuing to click.

Then the page rank of a page of a page (acc. to wikipedia) can be written as

$$r(v) = \frac{1-d}{N} + d(\sum_{w \in pa[v]} \frac{r(w)}{|ch[w]}) \tag{2}$$