Example Web Page

Tropical fish

From Wikipedia, the free encyclopedia

Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species. Fishkeepers often use the term *tropical fish* to refer only those requiring fresh water, with saltwater tropical fish referred to as *marine fish*.

Tropical fish are popular aquarium fish , due to their often bright coloration. In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Example Web Page

```
<html>
<head>
<meta name="keywords" content="Tropical fish, Airstone, Albinism, Algae eater,
Aquarium, Aquarium fish feeder, Aquarium furniture, Aquascaping, Bath treatment
(fishkeeping),Berlin Method, Biotope" />
…
<title>Tropical fish - Wikipedia, the free encyclopedia</title>
</head>
<body>
 …
<h1 class="firstHeading">Tropical fish</h1>
…
<p><b>Tropical fish</b> include <a href="/wiki/Fish" title="Fish">fish</a> found in <a
href="/wiki/Tropics" title="Tropics">tropical</a> environments around the world,
including both <a href="/wiki/Fresh_water" title="Fresh water">freshwater</a> and <a
href="/wiki/Sea_water" title="Sea water">salt water</a> species. <a
href="/wiki/Fishkeeping" title="Fishkeeping">Fishkeepers</a> often use the term
<i>tropical fish</i> to refer only those requiring fresh water, with saltwater tropical fish
referred to as <i><a href="/wiki/List_of_marine_aquarium_fish_species" title="List of
marine aquarium fish species">marine fish</a></i>.</p>
<p>Tropical fish are popular <a href="/wiki/Aquarium" title="Aquarium">aquarium</a>
fish , due to their often bright coloration. In freshwater fish, this coloration typically
derives from <a href="/wiki/Iridescence" title="Iridescence">iridescence</a>, while salt
water fish are generally <a href="/wiki/Pigment" title="Pigment">pigmented</a>.</p>
…
</body></html>
```

Link Analysis

- Links are a key component of the Web
- Important for navigation, but also for search
  - e.g., Example website
  - "Example website" is the anchor text
  - "http://example.com" is the destination link
  - both are used by search engines

Anchor Text

- Used as a description of the content of the *destination page*
  - *i.e., collection of anchor text in all links pointing to a page used as an additional text field*
- Anchor text tends to be short, descriptive, and similar to query text
- Retrieval experiments have shown that anchor text has significant impact on effectiveness for *some types of queries*
  - *i.e., more than PageRank*

Page Rank

- Billions of web pages, some more informative than others
- Links can be viewed as information about the *popularity(authority?) of a web page*
  - *can be used by ranking algorithm*
- *Inlinkcount could be used as simple measure*
- Link analysis algorithms like PageRank provide more reliable ratings
  - less susceptible to link spam

Random Surfer Model

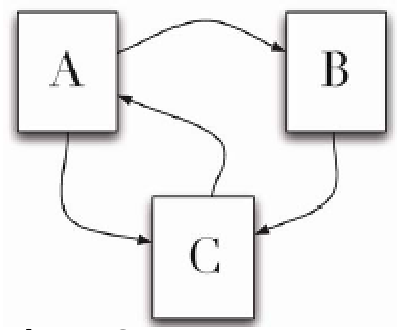
- Browse the Web using the following algorithm:
  - Choose a random number *r between 0 and 1*
  - If $r < \lambda$: *Go to a random page*
  - If $r \geq \lambda$: *Click a link at random on the current page*
  - Start again
- PageRank of a page is the probability that the "random surfer" will be looking at that page
  - links from popular pages will increase PageRank of pages they point to

Dangling Links

- Random jump prevents getting stuck on pages that
  - do not have links
  - contains only links that no longer point to other pages
  - have links forming a loop

- Links that point to the first two types of pages are called *dangling links*
  - *may also be links to pages that have not yet been crawled*

Page Rank



- PageRank (*PR) of page C = PR(A)/2 + PR(B)/1*
- More generally,

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

  – where B_u *is the set of pages that point to u, and* L_v *is the number of outgoing links from page v (not counting duplicate links)*

Page Rank

- Don’t know PageRank values at start
- Assume equal values (1/3 in this case), then iterate:
  - first iteration: *PR(C) = 0.33/2 + 0.33 = 0.5, PR(A) = 0.33, and PR(B) = 0.17*
  - second: *PR(C) = 0.33/2 + 0.17 = 0.33, PR(A) = 0.5, PR(B) = 0.17*
  - third: *PR(C) = 0.42, PR(A) = 0.33, PR(B) = 0.25*

- Converges to *PR(C) = 0.4, PR(A) = 0.4, and PR(B) = 0.2*

Page Rank

- Taking random page jump into account, 1/3 chance of going to any page when $r < \lambda$
- $PR(C) = \lambda/3 + (1 - \lambda) \cdot (PR(A)/2 + PR(B)/1)$
- More generally,

$$PR(u) = \frac{\lambda}{N} + (1 - \lambda). \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

  – where *N is the number of pages, λ typically 0.15*

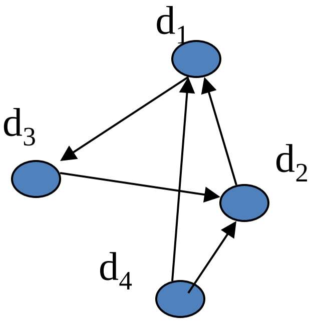
PageRank Algorithm

```
// P is the set of all pages; |P| = N
// S is the set of sink nodes, i.e., pages that have no out links
// M(p) is the set of pages that link to page p
// L(q) is the number of out-links from page q
// d is the PageRank damping/teleportation factor; use d = 0.85 as is typical

for each page p in P
        PR(p) = 1/N                          /* initial value */
while PageRank has not converged do
  sinkPR = 0
  for each page p in S                   /* calculate total sink PR */
    sinkPR += PR(p)

  for each page p in P
    newPR(p) = (1-d)/N                   /* teleportation */
    newPR(p) += d*sinkPR/N                 /* spread remaining sink PR evenly */
    for each page q in M(p)              /* pages pointing to p */
      newPR(p) += d*PR(q)/L(q)           /* add share of PageRank from in-links */

  for each page p
    PR(p) = newPR(p)
return PR
```

HITS: Capturing Authorities & Hubs

- Intuitions
  - Pages that are widely cited are good authorities
  - Pages that cite many other pages are good hubs
- The key idea of HITS (Hypertext-Induced Topic Search)
  - Good authorities are cited by good hubs
  - Good hubs point to good authorities
  - Iterative reinforcement…
- Many applications in graph/network analysis

The HITS Algorithm



$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

“Adjacency matrix”

Initial values: $a(d_i)=h(d_i)=1$

$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j)$$

$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j)$$

Iterate

Normalize:

$$\sum_i a(d_i)^2 = \sum_i h(d_i)^2 = 1$$

$$\vec{h} = A\vec{a}; \qquad \vec{a} = A^T \vec{h}$$

$$\vec{h} = AA^T \vec{h}; \quad \vec{a} = A^T A\vec{a}$$

Summary

- Link information is very useful
  - Anchor text
  - PageRank
  - HITS
- Both PageRank and HITS have many applications in analyzing other graphs or networks