Intelligent Data Analysis

PageRank

Peter Tiňo
School of Computer Science
University of Birmingham

Information Retrieval on the Web

Most scoring methods on the Web have been derived in the context of Information Retrieval:

- Use numerical vector representations $\mathbf{d}_j \in \mathcal{R}^T$ of documents $d_j \in \mathcal{D}$. Here, T is the number of terms.
- Apply some form of similarity measure in the vector space $\mathcal{R}^T$ of document representations.

But the Web is huge! For a given query, there may be unrealistically many "well-matching" documents.

Cannot rely just on term-based similarity between documents!

Need to rank pages/documents

In large hypertextual systems there is usually a strong topological interconnection structure.

Ideas that have already been around for some time

- Citation indexes in scientific literature
- Self-evaluating groups - each member of a group evaluates all the other members of the group

Idea: Rely on the democratic nature of the web!
Rank a page based on how it is embedded in the interconnection structure of the Web, not based on its content.

PageRank

Authority of a page p takes into an account:

- the number of incoming links (number of citations)
- authority of pages q that cite p with forward links
- selective citations from q are more "valuable" than flat (uniform) citations of a large number of pages.

Note the self-referencing nature of page authority!

Formalising PageRank

For a page $p \in \{1, 2, ..., N\}$ we define:

- $pa(p)$ - set of pages pointing to p
- h_p - number of hyperlinks from p (outdegree of p)
- "dumping factor" $0 < d < 1$ -
  d is the proportion of authority coming from other pages,
  $(1 - d)$ is the authority given to p by default ("for free")

PageRank equation (Brin, Page 1998):

$$x_p = (1 - d) + d \sum_{q \in pa(p)} \frac{x_q}{h_q}$$

PageRank in Matrix form

Stack all N page authorities x_p into a column vector
$\mathbf{x} = (x_1, x_2, ..., x_N)^T \in \mathcal{R}^N$.

Construct $N \times N$ transition matrix $\mathcal{W} = (w_{i,j})$:
$w_{i,j} = 1/h_j$, if the is a hyperlink from j to i;
$w_{i,j} = 0$, otherwise.

1_N - column N-dimensional vector of 1's.

$$\mathbf{x} = d \cdot \mathcal{W}\mathbf{x} + (1 - d) \cdot 1_N$$

PageRank - finding the solution

System of N equations with N unknowns x_p

$$\mathbf{x} = d \cdot \mathcal{W}\mathbf{x} + (1 - d) \cdot 1_N$$

The system is contractive and hence its dynamical form

$$\mathbf{x}(t) = d \cdot \mathcal{W}\mathbf{x}(t-1) + (1 - d) \cdot 1_N$$

converges (for any initial condition $\mathbf{x}(0)$) to a <u>unique</u> fixed point $\mathbf{x}_*$ (the solution):

$$\mathbf{x}_* = d \cdot \mathcal{W}\mathbf{x}_* + (1 - d) \cdot 1_N$$

Problem with dangling pages

Dangling pages - pages without hyperlinks.

If p is a dangling page, then p-th column of transition matrix $\mathcal{W}$ is null.
Each column corresponding to a non-dangling page sums to 1 (as in stochastic matrix).

Because of dangling pages, the transition matrix $\mathcal{W}$ is not stochastic. We cannot apply all the nice results available for (browsing) processes governed by stochastic matrices :-(

Getting around the problem of dangling pages

<u>Idea:</u> Introduce a "dummy" page r

- r has a link to itself
- every dangling page is made to point to r

Can you think what this does to the transition matrix $\mathcal{W}$?

<u>Hint:</u> We will get an <u>"extended" transition matrix $\bar{\mathcal{W}}$</u>

- Introduce a "dangling page indicator" vector $\mathbf{r} = (r_1, r_2, ..., r_N)$, where $r_p = 1$, if p is a dangling page, else $r_p = 0$.
- Stack $\mathbf{r}$ at the bottom of $\mathcal{W}$.
- To such row-extended matrix add an additional column of N 0's followed by a single 1.

Another way of getting around the problem of dangling pages

Idea: Make dangling pages point to all pages of the Web.

Construct matrix $\mathcal{V}$ with N equal rows $N^{-1}\mathbf{r}$:

$$\mathcal{V} = \frac{1}{N} 1_N \mathbf{r}$$

Modified PageRank equation (Ng et al. 2001)

$$\mathbf{x} = d \cdot (\mathcal{W} + \mathcal{V})\mathbf{x} + \frac{1-d}{N} \cdot 1_N$$

Solution $\tilde{\mathbf{x}}_*$ of the above system is related to the solution $\mathbf{x}_*$ of the original PageRank equation as follows:

$$\tilde{\mathbf{x}}_* = \frac{\mathbf{x}_*}{\|\mathbf{x}_*\|_1}$$

Recall that the L_1 norm $\|\mathbf{x}\|_1$ of $\mathbf{x}$ is

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + ... + |x_N|$$

Hence $\tilde{\mathbf{x}}_*$ can be thought of as representing probabilities of visiting nodes when surfing the web.

Think about stochastic interpretation of various forms of PageRank equations

Hints:

- Each page is a node in a graph.
- Nodes are connected as described by the hyperlink structure.
- Connections from a node p are weighted by probabilities of their usage (given that we are currently in p).
- The surfer never stops navigating.
- At each time step the surfer may become bored with probability $(1-d)$.
- When bored, the surfer jumps to any web page with uniform probability N^{-1}.

Web communities

Community - any subset of pages together with their hyperlink structure. Formally, it is a subgraph G of the Web graph.

Energy E_G of community G - sum of PageRank of all its pages. It is a measure of the community's authority.

$$E_G = \sum_{p \in G} x_p$$

Given a community G, we define 3 related communities:

- $out(G)$ - (sub)community of pages in G that point outside G
- $dp(G)$ - (sub)community of dangling pages in G
- $into(G)$ - external pages (outside G) that point to G

Community energy decomposition

- $|G|$ - number of pages in G. Larger communities tend to have higher authority.
- E_G^{into} - energy coming from outside G into G
- E_G^{out} - energy released from G to external pages
- E_G^{dp} - energy lost in dangling pages of G

The energy of G decomposes as (Bianchini et al., 2005):

$$E_G = |G| + E_G^{into} - E_G^{out} - E_G^{dp}$$

Community energy decomposition - cont'd

For any page p (inside or outside G), ρ_p is the fraction of all hyperlinks from p that point to G.

$$E_G^{into} = \frac{d}{1-d} \sum_{p \in into(G)} \rho_p x_p$$

$$E_G^{out} = \frac{d}{1-d} \sum_{p \in out(G)} (1-\rho_p) x_p$$

$$E_G^{dp} = \frac{d}{1-d} \sum_{p \in dp(G)} x_p$$

Lessons learnt

Community energy decomposition provides an insight into how the score migrates in the Web.

In order to maximise energy of G, we need to

- care about references received from other communities
- pay attention to links pointing outside G
- minimise the number of dangling pages inside G