PageRank and TrustRank algorithms

Group members:
Vinta Reethu (ES18BTECH11028)
Akash Tadwai (ES18BTECH11019)
Chaitanya Janakie (CS18BTECH11036)

Why Page Rank and Trust Rank:

To combat spamming in web pages. Web spam pages employ a variety of tactics to gain undeservedly high search engine rankings. For example, Thousands of keywords can be added to a website's home page, and the content is typically rendered invisible to humans using clever colour schemes. These pages will then be returned by a search engine, when queried with some of the keywords.

While humans can detect spam, manually evaluating a huge number of pages is too costly. Because it is impracticable and expensive to detect these spams manually, “**TrustRank**” was created to assist in making this work considerably faster and less expensive. It was introduced for the first time in the work **"Combating Web Spam with TrustRank**” by academics Zoltan Gyongyi and Hector Garcia-Molina of Stanford University and Jan Pedersen of Yahoo! in 2004. Today, major web search engines such as Yahoo! and Google uses this algorithm.

Web Model:

The web is represented as a graph G = (V,E) with a set V of N pages (vertices) and a set E of directed links (edges) connecting them. Two matrix representations of a web graph are introduced, both of which will play crucial roles in the coming sections. The “Transition matrix T” is one of them, and the other is “Inverse Transition Matrix U.” These are defined as follows.

$$\mathbf{T}(p,q) = \begin{cases} 0 & \text{if } (q,p) \notin \mathcal{E}, \\ 1/\omega(q) & \text{if } (q,p) \in \mathcal{E}. \end{cases}$$

$$\mathbf{U}(p,q) = \begin{cases} 0 & \text{if } (p,q) \notin \mathcal{E}, \\ 1/\iota(q) & \text{if } (p,q) \in \mathcal{E}. \end{cases}$$

Where w(q) is the no.of outlinks from node q or outdegree of q.
I(q) is the no.of inlinks to node q or indegree of q.

Page Rank Algorithm:

PageRank is a well-known algorithm that assigns global significance ratings to all web sites based on link information. The idea behind this method is that a web page is important if it is linked to by multiple other important web pages. It is based on mutual reinforcement between pages.
The PageRank score 'r' of a page 'p' is defined in matrix notation as:

$$\mathbf{r} = \alpha \cdot \mathbf{T} \cdot \mathbf{r} + (1-\alpha) \cdot \frac{1}{N} \cdot \mathbf{1}_N.$$

Where alpha is decay factor.

PageRank scores of pages can be computed iteratively until they converge or there is no significant change in the scores. The score of a web page p is made up of two parts: one portion comes from pages that link to p, and the other (static) part is the same for all web sites.

Oracle and Trust Functions:

An oracle function is a binary function 'O' formalises the idea of a person inspecting a page for spam. O(p) is 0 if the page is bad, 1 otherwise. However these invocations are time consuming and expensive,

So we'll depend on what we call the *approximate isolation* of the good set. This concept is self-evident, as bad pages are designed to deceive good ones. However the converse of this doesn't necessarily hold. The **Trust function** is provided to assess pages without relying on O. We define T as a trust function with a range of values between 0 (bad) and 1 (excellent) (good). Ideally it provides us the chance that p is good. Later, to relax the requirements, a Threshold has been introduced.

Initially the Trust function is defined as follows.

$$T_0(p) = \begin{cases} O(p) & \text{if } p \in \mathcal{S}, \\ 1/2 & \text{otherwise.} \end{cases}$$

Later, by taking advantage of the approximate isolation of good pages, the set S of L pages on which the oracle is invoked is still chosen at random. The Approximate Trust function is defined as

$$T_M(p) = \begin{cases} O(p) & \text{if } p \in \mathcal{S}, \\ 1 & \text{if } p \notin \mathcal{S} \text{ and } \exists q \in \mathcal{S}^+ : q \rightsquigarrow_M p, \\ 1/2 & \text{otherwise,} \end{cases}$$

Where we assign a score of 1 to all pages that are reachable from a page in S + in M or fewer steps.

Later it is observed that the trust has been reduced as we move farther from good ones. So to achieve the attenuation of trust, two methods have been proposed. First one is **Trust Dampening** and the second is **Trust Splitting**.

The Trust Rank Algorithm:

Step-1: We first evaluate the seed desirability of pages.
Step-2: Generate the corresponding rankings of the pages.
Step-3: Invoke the oracle function on L most desirable seed pages.
Step-4: Set the static score distribution vector d's elements that correspond to good seed pages to 1. And Normalize the vector d.

Step-5: Evaluate the Trust rank scores using biased page rank computation as follows.

$$\mathbf{t}^* = \mathbf{d}$$
$$\text{for } i = 1 \text{ to } M_B \text{ do}$$
$$\qquad \mathbf{t}^* = \alpha_B \cdot \mathbf{T} \cdot \mathbf{t}^* + (1 - \alpha_B) \cdot \mathbf{d}$$
$$\text{return } \mathbf{t}^*$$

* For selecting the seeds, we use the **Inverse PageRank Algorithm**.

PageRank vs TrustRank:

The PageRank algorithm does not include any information regarding a site's quality, and therefore does not overtly penalise poor behaviour. Indeed, we shall find that a site produced by a competent spammer receiving a high PageRank score is not rare. Whereas in TrustRank, on the other hand, is intended to distinguish between good and bad sites; we do not anticipate spam sites to have high TrustRank values.