PageRank – What Is Really Relevant in the World-Wide Web?

Ulrik Brandes and Gabi Dorfmüller

Universität Konstanz, Konstanz, Germany

No doubt, the most popular form of Internet usage is the World-Wide Web (WWW), a network of billions of files. It is made up, for the most part, of Web pages containing text and images that refer to each other via (hyper)links. Even if you spent your whole life, day and night, doing nothing but browsing pages, you’d see only a small fraction of the Web.\(^1\) To find something on the Web, it is therefore necessary to know where it is, or what links there.

Practically everyone surfing the Web therefore uses search engines, i.e., special pages on which the information sought is described using a few key words (query) to obtain a list of pages that may be relevant to the query (hits). Using many computer science methods, modern search engines are capable of organizing access to billions of Web pages, and scanning them for matches with a query within fractions of a second.

Since even a query term such as algorithm yields millions of hits, the results themselves are too large to be read completely. Search engines therefore sort their results in such a way that the seemingly most relevant hits are shown first.

Quiz:

How do search engines manage to find Web pages that seem relevant to us out of the millions matching a typical query?

As of today, the best known search engine is run by a company called Google,\(^2\) since it was the first search engine to not only sift through an enormous number of pages, but also to use a particularly clever algorithm to rank the results. Among, e.g., more than a million German-language hits matching Algorithmus, the site of the original project leading to this book currently ranks second only to the corresponding Wikipedia article.

\(^1\) Assume that you are spending a second per page – how many seconds does an average life last?

\(^2\) www.google.com
Besides many straightforward criteria such as the location of query terms in the page (in headlines? near each other?) and many unknown heuristic rules, a central element of the ranking strategy is an evaluation of the Web’s linking structure. This component is known as PageRank and is explained in this chapter.

Tourist Trails

Explanations of PageRank often use for motivation the idea that a page should be ranked as more relevant, the more frequently one would reach it on a random walk through the Web. We will explore this idea further, but with a completely different example. Imagine that, in the 18th century, mathematician Leonhard Euler had not proven the inexistence of, but rather found the long-sought tour crossing the seven bridges of Königsberg (cf. Chap. 28). This tour would be famous: It would be listed in all city guides, and tourists would walk the tour in droves. Of course, there would also be vendors selling souvenirs and refreshments in places that these tourists most frequently stroll by – but where are these places?

For a tour it does not matter where it is begun. Since, however, every bridge is crossed exactly once, we can at least be sure that every part is visited half as often as there are bridges leading there: One bridge is needed to get there, and another one to leave. The most promising selling spots are where the most bridges converge. In Königsberg, this would be Kneiphof (labeled A).

Alas, there is no such tour. So let’s assume tourists are wandering around with no particular goal or destination. More concretely, let them choose the next bridge to cross randomly and with equal probabilities (this is called “uniformly at random”) from all those that are feasible, including the one they just came across. How often do they arrive in a certain location?

The number $b$ of visits at, say, node $B$ can be described in terms of the number of immediately preceding visits at nodes that are connected to $B$, here $A$ and $D$. If the next bridge to continue with is chosen uniformly at random from all feasible bridges, we get from $A$ to $B$ in two out of five cases,