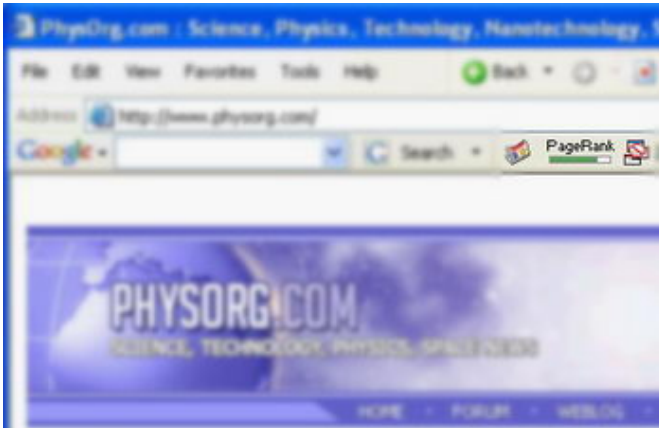


Prospecting for Scientific 'Gems' with Google



Science is all about quantitative measurement, so it should come as no surprise that scientists have a long tradition of measuring their influence on each other. Traditionally, the most important measure of scientific impact has been the number of citations an article receives -- but this method's chief virtue is its simplicity. The authors of any given paper are much more likely to reference recent works, so a result whose importance is not immediately recognized can end up with a much lower citation index than it deserves.

A well-written and relevant paper is typically referenced by two or three dozen papers, mostly from researchers working on the same highly specialized problem. A particularly useful or innovative result will often receive a hundred or more citations. Seminal works can achieve over a thousand citations, although they usually take decades to reach that point. While the system works well overall, it has no way of distinguishing a highly relevant citation from the polite mention of a colleague's work.

Many articles, for example, include an introduction section describing the history and current status of their specialized subject. This section can easily generate up to half of a paper's references, even though few of the results given mention are actually used.

A database of scientific literature is very similar in structure to the World Wide Web. Just as individual web pages are connected to each other by one-way links, journal articles are connected to each other by one-way citations. The number of external sites linking to a given website, or "in-degree", is equivalent to the citation index of a given article.

When Google tackled the problem of ranking websites by their influence, it didn't consider the in-degree to be an appropriate measure. This would make it too easy to inflate the importance of a site by creating a host of useless linking pages. Instead they crafted a customized statistic, the Google PageRank (GPR) algorithm.

To illustrate this algorithm, consider the webpage *PhysOrg.com*. PageRank finds every other webpage with a link to *PhysOrg.com*, and divides each neighbor's GPR by its total number of outgoing links. The GPR of *Physorg.com* is then calculated as the sum of all these factors. In other words, each site in the network can be thought of as evenly distributing its influence over all the sites that it links to. A page thus gains influence mainly by being associated with other influential pages. (The actual algorithm is a little more complicated, but this is its essential feature.) This method strikes a nice balance between content and connectivity, reducing the influence of high-traffic directories on the sites that they list.

So how can one calculate the GPR of any site when you first need to know the GPR of all its neighbors? It's not a problem to be solved by pencil and paper! The answer is found through a recursive calculation: every website in the network is initialized with the same GPR, so that a new GPR can be calculated for each website simultaneously. This calculation is repeated until all the values stabilize.

Researchers Patrick Chen and Sidney Redner at Boston University, along with their colleagues Huafeng Xie and Sergei Maslov at Brookhaven National Lab, recently applied the PageRank algorithm to all 353,268 articles published by the Physical Review between 1893 and 2003. It comes as no surprise that on average, GPR correlates nicely with the citation index. More interesting are the outliers—those articles that somehow achieve a high ranking with relatively few incoming references.

After applying PageRank, Chen et al. sorted the papers in this network by their GPR values. Their recent article provides a sampling of famous papers from the top hundred results. Number 85, with only three citations, is a startling poster child of this new approach! The paper in question is a classic example of delayed influence. While it was the first to present a model which today sees widespread use, its result was refined and popularized by other researchers in a separate article. The “child paper” has accumulated 680 citations but makes only ten references to other works itself. The original paper thus collects a large share of its child's impressive impact.

Nor is this the only example! Among the papers with over a hundred citations, most of the papers with an unusually high GPR are easily recognizable as seminal works. Such works compare favorably in overall influence with the very small population having over a thousand citations.

While “influence” may be easy to measure crudely, it is hard to measure reliably. These results show that although the two methods are comparable, Google's PageRank algorithm seems to identify important scientific papers more reliably than a simple citation index.

If there is a lesson here, it is this: in giving due credit, one should not be short-cited!

Reference: Patrick Chen, Huafeng Xie, Sergei Maslov, & Sidney Redner 2006, “Finding Scientific Gems with Google”, <http://xxx.lanl.gov/physics/0604130>

By Ben Mathiesen, Copyright 2006 PhysOrg.com

This document is subject to copyright. Apart from any fair dealing for the purpose of private study, research, no part may be reproduced without the written permission. The content is provided for information purposes only.