Print

PageRank

PageRank is another example of a collective inference procedure, which was developed by Page and Brin in 1999. It is the basis of Google's famous search engine. The PageRank value represents the probability of visiting a web page. It uses the democratic structure of the web as it is reflected by its link structure. In other words, the PageRank value of a web page A depends on the PageRank value of the pages that link to A and their number of outgoing links. The PageRank value of page A will be high if another page B refers to A, and B has a high PageRank value with only a few outgoing links. PageRank computation also takes into account random surfer behavior, whereby a surfer does not follow any links but randomly arrives on page A.

More specifically, the PageRank value of page A can be computed as PageRank A equals alpha times the sum, for i belonging to NA, of PageRank(i) divided by L(i) plus (1 minus alpha) times eA, where NA represents the pages that link to page A, PR(i) is the PageRank value of page i, L(i) is the number of outgoing links from page i, alpha is a damping factor representing the probability that a surfer will continue following the links on web pages, and thus (1 minus alpha) represents the probability of random surfing behavior. Alpha is usually set at approximately 0.85. Finally, eA represents the restart value for page A and is often assumed to be uniformly distributed. So for N pages, eA is usually set to 1 divided by N. We can now rewrite this expression in matrix notation as follows: r equals alpha times A times r plus (1 minus alpha) times e, where r is the vector of PageRank values, A is the column normalized adjacency matrix such that all columns sum to 1, and e is the restart vector. This equation will then typically be solved using an iterative procedure.

The PageRank algorithm can also be used to propagate behavior in any type of social network. Let's again consider the example of fraud detection. The adjacency matrix A now represents the people-to-people network. The known fraudulent nodes can be included in the restart vector e. In other words, the ith entry of the restart vector e equals 1 if the corresponding node is fraudulent, and 0 otherwise. It can then be normalized to make sure that all entries sum to 1. The r vector then contains the fraud ranking of each node. As the ranking goes higher, the node is increasingly influenced by fraud, compared to the fraud influence on the other nodes.

Here is the example fraud detection network that you saw earlier. The table shows the PageRank value for each node. The nodes with the highest PageRank values are nodes D, E, and F. This is not surprising because those nodes are known frauds. Node G also receives a high PageRank value or fraud score, because this node is directly influenced by the fraudulent nodes in its neighborhood. It is interesting to see that node G receives a higher PageRank value than the fraudulent node I.

Copyright © 2019 SAS Institute Inc., Cary, NC, USA. All rights reserved.

Close