Applied Artificial Intelligence
Project 3
Bayesian Network based reasoning using NETICA
Page Search
By Yoda

Description

This is a network that takes inspiration from Google's Page Ranking algorithm that displays the most relevant page to the user according to the keywords searched. The purpose of the project is to show how accumulation of factors can have an effect on which page the link is displayed and how that link will be relevant to the user.

Here we define relevance as how useful a particular page is to the user and how many times the user refers to this particular page.

For convenience we will address each webpage available on the internet as a node. There are numerous nodes available on the internet; it is up to the algorithm to sort out the best nodes from the rest of the nodes and to rank it accordingly. The factors which affect the rank and relevance of the node are following:

1. Keyword Score Percentile: When the user enters the keywords such as "NETICA software" the keywords "NETICA" is taken up as a keyword and every node is searched with respect to the keyword. Each node has a keyword score which depends on the following criteria:
   a. Keyword count of the page
   b. The number of hits on the site for any similar search such as "Netica models" etc.
   From here the keyword score percentile is calculated where every node's keyword score is tallied with the other nodes. Nodes with the highest keyword score is displayed on the top.

2. User profiling: Now days on the internet user profiling is very common. It is done for various reasons such as advertisements and as a method for the government to keep a track of the internet. User profiling can also be done by the search to make the search more accurate.
   For example, say if the user is a young male, it is very unlikely that he will be searching for a banking website.
   Thus, web search could be made for accurate using the user's age and gender.

3. Popular pages: For certain searches, there are limited number of pages that users like to visit. For example, for encyclopaedia related searches people are more like to visit Britannica and Wikipedia. For movie related searches people most likely visit IMDB and Rotten Tomato webpages. These are a collection of popular nodes that must on the top of the page display as the user is most likely to click on them.

4. Relevance of the node: This is also checked since it is important to know if the webpage displayed is safe for the user or not. This is down by checking the following criteria:
    a. Legal: It is checked if the page has generated any kind of spam hits or virus hits in the past. It is also checked if the page was blocked by the government.
    b. Keyword above cutoff score: A minimum cutoff keyword score is set and it is checked if the keyword cutoff score of the node is above the minimum cutoff score or not. The purpose behind this step is that below a certain minimum cutoff score the page ceases to be relevant to the user.

From the above factors the relevance of the node and the chances of node being displayed on a page is calculated.

From the Page Rank of the node, the overall relevance of the node to the user is calculated. The more relevant the nodes are, the more efficient our page search algorithm is.

Test Cases

1. Set the Keyword Search to True and the Keyword score to VH and observe how the Decide Page Rank is decided.
2. Change Spam_Hits, Virus_Hits and other factors and observe how Legal Node changes.