Trust Rank using PREGEL Framework

Vijay Varma K AI20BTECH11012
Sachin Karumachi AI20BTECH11013
Abhiroop Ch AI20BTECH11005
Jeevan S CS20BTECH11047
Jatin Kumar CS19BTECH11036

May 4, 2023

1 INTRODUCTION

TrustRank is a link analysis algorithm that aims to identify trustworthy web pages and demote spam or low-quality pages in search engine results. It is a variant of the PageRank algorithm, which ranks web pages based on the number and quality of incoming links.
In the TrustRank algorithm, a set of seed pages that are considered trustworthy are manually selected. These seed pages are then used to propagate trust scores to other pages in the graph using a modified version of the PageRank algorithm. The modified algorithm ignores links from untrustworthy pages and assigns higher weight to links from trustworthy pages.
The resulting trust scores can be used to rank web pages in search engine results or to identify spam pages that have been artificially inflated through link farms or other black hat techniques. TrustRank has been used by major search engines, such as Google and Yahoo, to improve the quality of their search results and combat web spam.

2 PROBLEM STATEMENT

We want to implement Trust rank on Iron dealers data provided using the Pregel framework to find out bad dealers.

3 MOTIVATION

Running Analytics on huge graph data is a complex task and very time consuming and one of the things we tend to optimize is Trust rank algorithm which aims to identify trustworthy pages on the web(Think web like a Huge Graph), so we use Pregel framework implementing the trust rank algorithm using distributed computing framework to reduce the time and manage resources efficiently.

4 DATASET

We have two datasets. In the first dataset,

- The first Dataset consists of 3 columns “Seller ID”, “Buyer ID” and “Value”
  - *Seller ID*: Unique ID given to a dealer, this column consists of dealers selling goods (This is of type INT).
  - *Buyer ID*: Unique ID given to a dealer, this column consists of dealers buying goods from seller (This is of type INT).
  - *Value*: It indicates the value of transaction between seller and buyer (This is of type FLOAT).

- The second dataset we have 'Bad Id' which are the bad nodes in our first dataset
    - *Bad ID*: Unique ID given to a dealer, this column consists of dealers who are known to be doing illegal activities (This is of type INT).

5 TRUSTRANK ALGORITHM

1. *Select Trusted Seed Pages*: The first step in the TrustRank algorithm is to manually identify a set of trusted seed pages that are known to be reliable and trustworthy. These pages are usually selected based on human expertise or other signals of trustworthiness, such as government websites or educational institutions.

2. *Compute Trust Scores*: In the second step, the trust scores of the seed pages are set to 1, and the trust scores of all other pages in the graph are set to 0. The modified PageRank algorithm is then used to propagate the trust scores from the seed pages to the other pages in the graph.

3. *Propagate Trust Scores*: During each iteration of the PageRank algorithm, each page computes its trust score based on the trust scores of its incoming links. However, in the TrustRank algorithm, links from untrustworthy pages are ignored, and links from trustworthy pages are given higher weight. This ensures that trust scores are only propagated along a path of trustworthy links.

4. *Rank Pages by Trust Score*: Once the trust scores have been computed, the pages in the graph can be ranked according to their trust score. Pages with higher trust scores are considered to be more trustworthy, and pages with lower trust scores are considered to be less trustworthy.

5. *Demote Low-Quality Pages*: Finally, the trust scores can be used to demote spam or low-quality pages in search engine results. Pages with low trust scores are typically given lower rankings or excluded from search results altogether, while pages with high trust scores are given higher rankings and more visibility.

6 DRAWBACKS OF THE TRUSTRANK ALGORITHM

1. *Reliance on Manual Seed Selection*: The TrustRank algorithm relies on the manual selection of seed pages that are assumed to be trustworthy. This can introduce biases and subjectivity into the algorithm, and it can also limit the effectiveness of the algorithm if the seed pages are not representative of the entire web.

2. *Difficulty in Defining Trustworthiness*: The concept of trustworthiness is difficult to define and can vary depending on the context. For example, a page that is trustworthy for a medical topic may not be trustworthy for a political topic. This can make it challenging to identify trustworthy seed pages and to develop a robust definition of trustworthiness.

3. *Vulnerability to Manipulation*: Like other link-based algorithms, the TrustRank algorithm can be vulnerable to manipulation by webmasters who seek to boost their page rankings. This can be done by creating artificial links to trustworthy seed pages or by using other black hat SEO techniques.

4. *Limited Scope*: The TrustRank algorithm is designed specifically for identifying trustworthy pages in web search results. It may not be as useful for other graph-based applications, such as social network analysis or recommendation systems.

7 ALGORITHM IMPLEMENTATION

Since we are using Bad nodes instead of good nodes so we are assuming that bad pages only promote bad pages(instead of good pages promoting good pages).
We have updated the pagerank algorithm that has been given in the Pregel framework.
From given csv file we have generated a two text files, one consisting of all nodes and other

consisting of edges as specified in the page rank implemented in Pregel framework.
Changes that are done are:

1. Initializing all the bad nodes to '1/(number of bad nodes)' and all other nodes to '0'
2. Updating the "updateOutGoingVertices" function such that created a dict where all nodes cited by "src node" are in dictionary as a list with "src node" as key
3. There is a error when there are no outgoing edges we have updated it by taking 1 where there are no outgoing links
4. Hyperparameters taken:
   (a) $dampingFactor = 0.85$
   (b) $iterations = 50$

This gives output of the trustrank algorithm getting list of nodes such that trust score of pages are in decreasing order, since we have taken initial condition that we are taking Bad nodes as seed nodes and assuming that bad pages promote bad pages we get list of nodes such that bad trust score of pages are in decreasing order(All the bad pages have high bad trust score).

8 RESULTS

Once we finished implementing the code, we got the trust ranks of the nodes good and bad in an array.

Trust ranks of given bad nodes, good nodes with least trust score, bad nodes with highest trust score

Sno	Given Bad Nodes	Trust Score	Good Nodes	Trust Score	Bad Nodes	Trust Score
1	1309	0.0064	1009	0.0002	1144	0.0212
2	1259	0.0002	1012	0.0002	1007	0.0209
3	1568	0.0002	1045	0.0002	1088	0.0189
4	1147	0.0020	1047	0.0002	1201	0.0137
5	1393	0.0002	1051	0.0002	1173	0.0107
6	1039	0.0038	1066	0.0002	1041	0.0093
7	1210	0.0023	1078	0.0002	1043	0.0089
8	1042	0.0016	1081	0.0002	1094	0.0087
9	1045	0.0002	1082	0.0002	1381	0.0086
10	1256	0.0002	1090	0.0002	1122	0.0077
11	1668	0.0002	1106	0.0002	1330	0.0073
12	1163	0.0005	1115	0.0002	1138	0.0072
13	1007	0.0209	1132	0.0002	1309	0.0064
14	1034	0.0041	1134	0.0002	1105	0.0064
15	1832	0.0002	1135	0.0002	1033	0.0061
16	1099	0.0017	1142	0.0002	1283	0.0058
17	1488	0.0005	1145	0.0002	1246	0.0057
18	1801	0.0002	1146	0.0002	1049	0.0057
19	1076	0.0054	1151	0.0002	1076	0.0054
20	1944	0.0002	1151	0.0002	1050	0.0053

9 CONCLUSION

- The TrustRank algorithm is a modification of the PageRank algorithm that aims to identify and rank trustworthy web pages. By relying on a set of manually selected trusted seed pages and propagating trust scores along a path of trustworthy links, the algorithm can improve the quality of search engine results by demoting spam or low-quality pages.
- The TrustRank algorithm has some limitations and potential drawbacks, including reliance on manual seed selection, difficulty in defining trustworthiness, vulnerability to manipulation, and limited scope.

- Despite these limitations, the TrustRank algorithm can be a valuable tool for improving the quality of web search results, and it highlights the importance of trust and credibility in the modern digital age.

10 REFERENCES

- ChatGpt
- TrustRank algorithm paper -link
- PageRank algorithm paper -link