


Práctica 2.3

Introducción

PageRank es una marca registrada y patentada por Google el 9 de enero de 1999 que ampara una familia de algoritmos utilizados para asignar de forma numérica la relevancia de los documentos (o páginas web) indexados por un motor de búsqueda.

Modelo matemático



PAGE RANK

- Donde:
- **PR(A)**, es el Page Rank de la página A
- **d**, es un factor de amortiguación [0,1]
- **PR(i)**, son los valores de PR que tienen cada una de las páginas *i* que enlazan a A
- **C(i)**, es el número total de enlaces salientes de la página *i* (sean o no hacia A)

$$PR(A) = (1 - d) + d \sum_{i=1}^n \frac{PR(i)}{C(i)}$$

Desarrollo

Al implementar el algoritmo (se muestra a continuación), tomé el factor de amortiguación default, es decir, $d = 0.85$. Tras evaluar los archivos CSV de Oracle (nodos y aristas) con el código, obtuve los siguientes resultados:

```

yorchpave@YorchPC:~/Dropbox/ITESM/Bases de Datos Avanzadas/Prácticas/Práctica 2.3$ python3 main.py
PageRank Results:

Oracle          3.780261323660153
Oracle OpenWorld      2.1403222898215013
Oracle Cloud          1.3240182132690874
Oracle IT Infrastructure 1.2502254973895397
Oracle University     1.232313764717813
Oracle Developers     1.2145616366642829
ORACLE TEAM USA       1.1831565339194872
Oracle Financial Services 0.9260864122523074
Sean D. Tucker        0.8943324950616413
I <3 Java             0.8514935807408974
Oracle Management Cloud 0.7766725994914256
Oracle Solaris         0.771115828056163
Oracle Database        0.7703379351786628
Oracle Magazine        0.7545829737695691
OracleCRM              0.7066008631953301
Oracle Linux           0.7024217501093857
Oracle System Developers 0.6998843884019434
Oracle Customer Experience 0.6750019088598783
OracleHCM              0.6739953146297413
Oracle PartnerNetwork 0.6170520808836889
Oracle Midsize         0.6051265379453872
Oracle Code One        0.5982863432596153
MySQL                 0.5943769042066149
Oracle Applications    0.5903043707700777
The Oracle ACE Program 0.5750994099427096

```

```

Oracle Retail          0.575007926438752
Oracle Business Analytics 0.5425210991002047
Oracle Profit Magazine Online 0.5074382682160954
James Spithill         0.505425303047784
Cloud Odyssey: A Hero's Quest 0.5043034530774868
Oracle Communications    0.489260049243054
Explore Oracle          0.4718877722272383
Oracle Latinoamérica     0.4324907470265004
Oracle Supply Chain Management 0.4310141499929051
Oracle Product Lifecycle Management 0.4204177655576716
Oracle Security         0.4131454974992364
EAA - Young Eagles      0.4056216169375606
Oracle Developer Community: ArchBeat Page 0.4034730823401585
Oracle Learning Library (OLL) 0.3916635549611638
Oracle ERP Cloud         0.3600385877010015
Oracle Health Sciences    0.34649424656608785
Oracle Primavera Enterprise Project Portfolio Management 0.31040016627429845
High Ground            0.29137904283480665
McGraw-Hill Education    0.26878435451872507
IOUG - Independent Oracle Users Group 0.257827153173969
Be The Match            0.21557596173696186
International Federation of Red Cross and Red Crescent Societies 0.21557596173696186
Second Harvest Food Bank 0.21557596173696186
Iron Man                0.21557596173696186
BNP Paribas Open        0.21557596173696186
yorchpave@YorchPC:~/Dropbox/ITESM/Bases de Datos Avanzadas/Prácticas/Práctica 2.3$ █

```

Código:

Graph.py

class Node(object):

```

    def __init__(self, id, name, pageRank = 0):
        self.id = id
        self.name = name
        self.adj = {}
        self.invAdj = {}
        self.connected = 0
        self.pageRank = pageRank

```

```

def addConnection(self, node, cost):
    self.adj[node.id] = [cost, node]
    node.invAdj[self.id] = self
    self.connected += 1

def updatePageRank(self, d = 0.85):
    accum = 0
    for node in self.invAdj.values():
        accum += node.pageRank / node.connected
    self.pageRank = (1 - d) + (d) * accum

```

```

class Graph (object):
    def __init__(self, name="Graph"):
        self.name = name
        self.size = 0
        self.nodes = {}

    def insertNode(self, node):
        self.nodes[node.id] = node
        self.size += 1

    def pageRank(self, iterations = 0):
        keys = self.nodes.keys()
        for _ in range(iterations):
            for key in keys:
                self.nodes[key].updatePageRank(d=0.85)
            pageRankDictionary = {}
            for key in keys:
                pageRankDictionary[self.nodes[key].name] = self.nodes[key].pageRank
            keyValue = zip(pageRankDictionary.keys(), pageRankDictionary.values())
            return dict(sorted(keyValue, key = lambda x: x[1], reverse=True))

    def fromCSV(self, pathToNodes, pathToEdges):
        import pandas as pd
        csvNodes = pd.read_csv(pathToNodes)
        for i, row in csvNodes.iterrows():
            self.insertNode(Node(row['id'], row['label']))
        csvEdges = pd.read_csv(pathToEdges)
        for i, row in csvEdges.iterrows():
            self.nodes[row['Source']].addConnection(self.nodes[row['Target']], 1)

```

main.py

```

from Graph import *

if __name__ == '__main__':
    myGraph = Graph("Pages")
    myGraph.fromCSV("Oracle_Nodes.csv", "Oracle_Edges.csv")

    print("PageRank Results: \n")

```

```
pageRank = myGraph.pageRank(50)
for key, value in zip(pageRank.keys(), pageRank.values()):
    print(key + "\t\t" + str(value))
```

```
del myGraph
```

¿Cuál es el nodo más influyente?

Como podemos ver, Oracle es el nodo más influyente con un pagerank $PR = 3.78$. Esto quiere decir que es el nodo más importante del grafo porque Oracle apunta a cierto número de páginas importantes y, al mismo tiempo, páginas importantes o relevantes apuntan a Oracle.

¿Cuál es el nodo menos influyente?

Se puede decir que existe un quíntuple empate entre las páginas:

1. Be The Match
2. International Federation of Red Cross and Red Crescent Societies
3. Second Harvest Food Bank
4. Iron Man
5. BNP Paribas Open

con un PageRank $PR = 0.21$. Estas fueron las páginas menos importantes ya que, a pesar de apuntar a páginas relevantes, estas no son apuntadas de vuelta por páginas relevantes.

Conclusiones

Analizando los resultados (imágenes anexas), me llamó la atención que los nodos más influyentes fueran de la misma empresa de Oracle. Aunque por una parte, esto es de esperarse ya que Oracle es una empresa altamente reconocida a nivel Mundial, por lo que muchas páginas relevantes apuntan a las diferentes páginas de Oracle. Mientras que, los nodos menos influyentes, no tienen nada que ver con Oracle y no son apuntadas por páginas relevantes, razón por lo cual su influencia es mínima.

Referencias

<https://es.wikipedia.org/wiki/PageRank>

PAGE RANK M. en C. Rodolfo Rubén Álvarez González