МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

РАЗРАБОТКА ИНСТРУМЕНТАЛЬНЫХ СРЕДСТВ РАНЖИРОВАНИЯ ВЕРШИН СЛОЖНЫХ СЕТЕЙ И ИХ ПРИМЕНЕНИЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Айрапетяна Оганеса Владимировича

Научный руководитель
доцент, д. ф.-м. н. _____ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2023

ВВЕДЕНИЕ

Актуальность темы работы связана с тем, что методы ранжирования узлов сложных систем активно используются в различных практических задачах. В частности, подобные алгоритмы применяются при ссылочном ранжировании web-страниц для поиска нужной информации по запросу пользователей, для нахождения важных узлов (хабов) сложных сетей в задачах, связанных с анализом социальных сетей, а также анализом диффузии информации и распространением инфекций в сетевых структурах. В данной работе будет изучаться сеть, узлами которой являются страны, а дугами - объемы международной торговли между каждыми двумя странами. Анализ такого рода сетей, описывающих межстрановые потоки торговых объемов, представляется достаточно важным, так как от стабильной международной торговли зависит состояние мировой экономики и участвующих в ней стран, что в свою очередь напрямую влияет на благополучия граждан. В связи с этим в данной работе с использованием методов ранжирования будет проведен сравнительный анализ сетей международной торговли в допандемийный 2019 год, в пандемийный 2020 год и в послепандемийный 2021 год, что также особенно интересно в контексте нынешней сложной экономико-политической ситуации в мире.

Цель бакалаврской работы изучение методов ссылочного ранжирования, применение их к анализу влияния стран в международной торговле за 2019, 2020 и 2021 года и сравнения их.

Объект исследования межстрановые потоки торговых объемов.

Предмет исследования выявление наиболее влиятельных стран мира в межстранновой торговле.

Для достижения поставленных целей в работе необходимо решить следующие **задачи:**

— определить основные понятия необходимые для описания графов;
— определить понятие взвешенного орграфа;
— определить основные понятия алгоритмов ссылочного ранжирования;
— определить основные понятия алгоритма `PageRank`;
— определить понятие цепи Маркова;

— обосновать теорему о Марковской цепи;
— определить понятие спектральная щель;
— определить понятие сжимающегося отображения;
— вычислить уравнение из которого необходимо будет искать вектор `PageRank`;
— обосновать эффективность алгоритма;
— рассмотреть пример работы `PageRank`;
— доказать корректность алгоритма с математической точки зрения;
— реализовать алгоритм `PageRank`;
— определить основные понятия алгоритма `Eigenvector Centrality`;
— определить с нормализацией вектора;
— доказать то, что алгоритм сходится;
— рассмотреть пример работы `Eigenvector Centrality`;
— реализовать алгоритм `Eigenvector Centrality`;
— определить основные понятия алгоритма `Betweenness Centrality`;
— определить основные понятия алгоритма Дейкстры;
— рассмотреть пример работы алгоритма Дейкстры;
— доказать корректность работы алгоритма Дейкстры;
— реализовать алгоритм Дейкстры;
— рассмотреть пример работы `Betweenness Centrality`;
— реализовать алгоритм `Betweenness Centrality`;
— преобразовать данные о межстрановых потоках торговых объемов в граф;
— применить к полученному графу ранее реализованные алгоритмы;
— получить, интерпретировать и сравнить результаты;
— сформулировать выводы о проделанной практической части работы.

Практическая значимость изучение алгоритмов ссылочного ранжирования позволит выявлять главные (центральные) компоненты/объекты в системах, которые можно представить в виде графа.

Структура и содержание бакалаврской работы. Работа состоит из введения, трех разделов, заключения, списка используемых источников, содержащий более 20-и наименований, и трех приложений. Общий объем работы составляет более 40 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введение** обосновывается актуальность темы работы, формируется цели и задачи, отмечается практическая значимость полученных результатов.

В **первом** разделе приведены основные понятия теории графов.

Во **втором** разделе рассмотрены три простейших алгоритма ссылочного ранжирования:

— Импакт-фактор,

— `Hilltop`,

— `BrowseRank`;

и три более сложных алгоритма:

— `PageRank`;

— `Eigenvector Centrality`;

— `Betweenness Centrality`.

Дается определение `PageRank`, принципы его работы, где может приниматься. Водятся необходимые значения и описывается общий случай выражения $PR(\mu)$:

$$PR(\mu) = \sum_{\upsilon \in B_\mu} \frac{PR(\upsilon)}{L(\upsilon)}.$$

Разъясняется то, что значение $PR(\mu)$ зависит от значений каждой страницы $PR(\upsilon)$, содержащихся в наборе B_μ, деленных на количество ссылок со страницы υ.

Рассматривается ориентированный взвешенный граф. Граф имеет n вершин. Каждой паре вершин соответствует некоторый вес $p_{i,j} \geq 0$. Ребро, выходящее из вершины i в вершину j, имеет вес $p_{i,j} > 0$. Если из вершины i в вершину j ребра нет, то полагаем $p_{i,j} = 0$. Число $p_{i,j}$ интерпретируется как вероятность перейти из вершины i в вершину j. Набор чисел $p_{i,j}, i, j = 1, 2, \ldots, n$ удобно будет записать в виде матрицы

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix}.$$

Поскольку распределение вероятностей должно быть нормировано на единицу, то для любой вершины i имеет место равенство

$$\sum_{j=1}^{n} p_{i,j} = 1,$$

и матрица P является стохастической по строкам.

Дается определение Марковской цепи, приводится теорема о Марковской цепи и абстрактный пример с пояснениями.

Водится предложение о «центральной» web-страницы.

Приводится более строгая формула для вычислений значимости узлов алгоритмом `PageRank`:

$$\sum_{i=1}^{n} \upsilon_i p_{i,j} = \upsilon_j$$

или в векторном виде $\upsilon^T P = \upsilon^T$, имеющему единственное решение, удовлетворяющее $\sum_{i=1}^{n} \upsilon_i = 1$.

Поясняется почему предположение о «центральной» web-страницы обеспечивает единственность υ.

Приводится формула Колмогорова–Чепмена:

$$p^T(t+1) = p^T(t)P.$$

Отмечается, что приведенные выше рассуждения справедливы в предположении существования предела $\lim\limits_{t \longrightarrow \infty} p(t) = \upsilon$. Казалось бы, что предположения о «центральной » web-странице будет достаточно и тут. Однако, как показывает простейший пример на рисунке 1, в котором $n = 2$, $p_{1,1} = p_{2,2} = 0$, $p_{1,2} = p_{2,1} = 1$, хотя вектор $\upsilon = (\frac{1}{2}, \frac{1}{2})^T$ существует и единственен, предел $\lim\limits_{t \longrightarrow \infty} P(t)$ не существует, поскольку с ростом t будет происходить периодическое чередование нулей и единиц в каждой компоненте вектора $p(t)$.

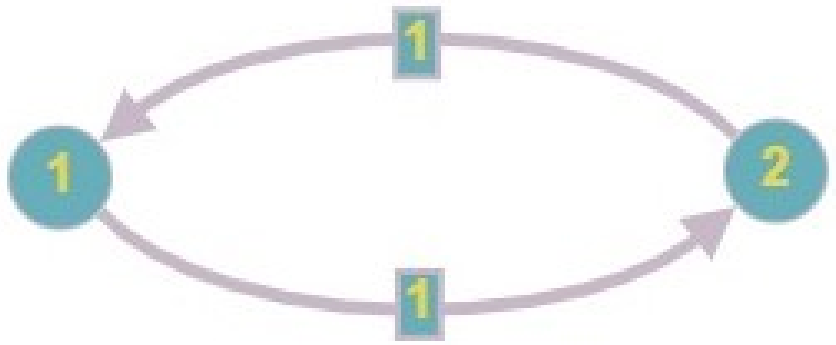

Рисунок 1 – Периодическая марковская цепь с периодом 2

Обосновывается то, что если выполняется условие «непериодичности», то предел $\lim\limits_{t \longrightarrow \infty} p(t) = \upsilon$ действительно существует. Более того, эти два условия (существования «центральной» web-страницы и «непериодичности») являются не только достаточными, но и необходимыми для существования предела. Описывается, в чем заключается условие «непериодичности». Из «центральной» web-страницы выходит много различных гиперссылок, которые в конце снова приводят на «центральную» web-страницу. Условие «непериодичности» означает, что наибольший общий делитель последовательности длин всевозможных маршрутов (начинающихся и заканчивающихся на «центральной» web-странице) равен 1. Уточним, что длина маршрута равна числу ребер, вошедших в маршрут. В типичных web-графах оба отмеченных условия выполняются, поэтому в дальнейшем эти два условия будут считаться выполненными..

Вычисляется уравнение из которого и необходимо искать вектор `PageRank`, с помощью двух предположений.

Предположим сначала, что для рассматриваемого web-графа существует такая web-страница, на которую есть ссылка из любой web-страницы, в том числе из самой себя, более того, предположим, что на каждой такой ссылке стоит вероятность, не меньшая, чем γ. Для такого web-графа имеет место неравенство $\alpha \geq \gamma$.

Предположим далее, что в модели блуждания по web-графу имеется «телепортация»: с вероятностью $1 - \delta$ «человек действует как в исходной модели», а с вероятностью δ «забывает про все правила» и случайно равновероятно выбирает среди n вершин одну, в которую и переходит. Тогда, если

ввести квадратную матрицу E размера n на n, состоящую из одинаковых элементов $\frac{1}{n}$, уравнение $p(t+1) = P^T p(t)$ примет вид

$$p(t+1) = ((1-\delta)P^T + \delta E)p(t).$$

В таком случае вектор `PageRank` необходимо будет искать из уравнения

$$\upsilon = ((1-\delta)P^T + \delta E)\upsilon.$$

При $0 < \delta < 1$ уравнение вышеприведенное уравнение гарантированно имеет единственное в классе распределений вероятностей решение. Более того, для спектральной щели матрицы $(1-\delta)P + \delta E$ имеет место оценка $\alpha \geq \delta$.

Обосновывается эффективность алгоритма. Такой алгоритм способен ранжировать все web-страницы во всей всемирной паутины менее чем за сутки, однако при условии, что найдется железо которое будет способно справиться со столь массивными данными иными словами мощности и памяти будет недостаточно.

Продемонстрирован пример работы `PageRank` и приведена реализация алгоритма.

```
U_past_has_alpha = []
while True:
    U = 0.8 * (np.dot(P, U)) + 0.2 * U0
    if str(U) == str(U_past_has_alpha):
        break
    U_past_has_alpha = U
print(U)
```

Дается определение алгоритму центральность по собственному вектору.

Водятся необходимые значения и описывается общий случай вычисления центральность вершины μ:

$$x_\mu = \frac{1}{\lambda} \sum_{t \in B_\mu} x_t$$

или, в общем в виде

$$x_\mu = \frac{1}{\lambda} \sum_{t \in V} p_{\mu,t} x_t.$$

Объясняется, что в общем, будет много разных собственных λ значений, для которых существует ненулевое решение по собственным векторам. Собственный вектор определяется только с точностью до общего множителя, поэтому четко определены только отношения центральностей вершин. Чтобы определить абсолютную оценку, необходимо нормализовать собственный вектор, например, так, чтобы сумма по всем вершинам была равна 1 или общему числу вершин n.

Обосновывается выбор значения λ для нормализации. Необходимо подбирать значение λ на каждой итерации таким образом, что сумма значений собственного вектора равнялось 1 будет неудобно, так как в предстоящей задаче количества вершин будет 100+, за счет этого значения собственного вектора будут близки к нули, различаться между собой, в общем случаи они будут лишь 3-ым, а то и 4-ы знаком после запятой. Из этих соображений значение λ будет подбираться таким образом, что сумма значений собственного вектора равнялось n.

Приводится пример работы и реализация алгоритма `Eigenvector Centrality`.

```
def X_Fun(X, n, P):
    X0 = [0] * n
    for i in range(n):
        for j in range(n):
            X0[i] += P[i, j] * x[j]
    for i in range(n):
        x[i] = X0[i]
    return x
X0 = np.zeros([n])
while (X != X0).all():
    for i in range(n):
        X0[i] = X[i]
    lamda = sum(X) / n
    X = X_Fun(X, n, P)
    X = [i / lamda for i in X]
```

Дается определение `Betweenness Centrality`, принципы его работы, где может приниматься.

Приводятся принципиальные различия между `Betweenness Centrality` и вышеперечисленным алгоритмами.

Водятся необходимые значения и описывается алгоритм Дейкстры.

Доказывается корректность алгоритма Дейкстры.

Приводится пример работы алгоритма Дейкстры.

Приводится пример работы алгоритма `Betweenness Centrality`.

В **третьем** разделе формулируется главная задача бакалаврской работы, она состоит в анализе влияния стран на основе данных о международной торговле за 2019–2021 годы, а также в сравнении и интерпретации полученных результатов. Основываясь на данных об объеме международной торговли стран и применением трех популярных алгоритмов для ранжирования вершин графов, сделать вывод о степени важности страны в международной торговле.

Проводится нормализация данных. Дело в том, что веса ребер слишком большие. При ранжировании достаточно массивного взвешенного графа, вес ребра которого является шести-семизначными числами, в результате будут получены веса вершин графа порядка $1e1000000$. Так как с такими числами сложно работать, необходимо нормализовать данные. Нормализация будет проводится по формуле '`MinMax`':

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

Совершается преобразование данных о межстрановых потоковых торговых объемов в граф.

Описывается матрица смежности. На главной диагонали все элементы нулевые, это означает, что ни одна из стран не экспортирует и не импортирует ничего в саму и для самой себя. Элементы (i, j), $i > j$, означают, что страна j получает денежные средства от страны i за импорт. Элементы (i, j), $i < j$, означают, что страна i получает денежные средства от страны j за импорт.

Применяются ранее реализованные алгоритмы к графу, представляющие собой межстрановые потоки торговых объемов.

Объясняется особенность интерпретация значения отдельной взятой вершины. Прежде чем приступить к интерпретации результатов, нужно понять то, что отдельно взятое значение полученное в векторе результат, само по

себе не о чем не говорит. К примеру, нельзя по значению PR сказать большое оно или нет. Единственное, что можно, это примерно оценить значимость вершины и то только в том случаи если на вход подавалась матрица сумма элементов которых в каждом столбце была константой. Если сумма элементов в каждом столбце равны друг другу, то можно оценить значимость вершины по его PR. Но в данном случаи сумма элементов в столбцах матрицы не равны и не могли были быть равны. Когда одна web-страница ссылается на другую, в матрице соответствующий элемент принимает значение

$\frac{1}{x}$, где x – количество страниц на которую ссылается страница.

В данном случаи когда одна страна экспортирует или импортирует товар в другую, то соответствующие значение в матрице будет на прямую зависеть от цены товаров которые экспортировали или импортировали. Конечно можно было бы поделить все элемент в столбцах на сумму всех элементов соответствующего столбца, но это будет не правильно, так как все элементы уменьшаться в разное число раз, что может серьезно повлиять на конечный результат.

Интерпретируются полученные результаты, проводится сравнение между результатами полученных от каждого алгоритма.

Формулируется выводы о проделанной практической части работы.

В **заключении** приведены основные результаты бакалаврской работы.

ОСНОВНОЕ РЕЗУЛЬТАТЫ

1. Определены основные понятия, необходимые для описания взвешенных орграфов.

2. Определены основные понятия, связанные с алгоритмами ссылочного ранжирования. Описаны и реализованы алгоритмы:

— `PageRank`;

— `Eigenvector Centrality`;

— `Betweenness Centrality`.

Рассмотрены примеры работ алгоритмов. Приведена их реализация в **приложениях A, B и C** соответственно.

3. Получены результаты по ранжированию данных о межстрановых потоковых торговых объемов за период 2019–2021 годы с помощью ранее реализованных алгоритмов. Результаты интерпретированы. Сравнение результатов проведено.

4. Сделан выводы о практической части работы и работы в целом.