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 ПРИЛОЖЕНИЯ ДЛЯ АНАЛИЗА УСТОЙЧИВОСТИ БИОСИСТЕМ С ИСПОЛЬЗОВАНИЕМ ГРАФОВЫХ МОДЕЛЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 452 группы
направления 38.03.05 — Бизнес-информатика
механико-математического факультета
Землянской Марии Юрьевны

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

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

Саратов 2019

ВВЕДЕНИЕ

Данная работа посвящена разработке приложения для анализа устойчивости биосистем с использованием графовых моделей. В настоящее время важным аспектом существования цивилизованного общества является сохранение экологии и поддержание равновесия различных экосистем, в частности, демэкологии (экология популяций). Любой вид растений, животных, микроорганизмов представлен своеобразной группировкой организмов, включающей особи со сходной морфологией, физиологией, поведением и общим генофондом. Каждый вид занимает определенный ареал, в разных частях которого наблюдаются разные условия существования.

Группировки особей внутри вида называются популяциями. На уровне популяций происходят основные адаптации, естественный отбор и эволюционные процессы. Разнообразие популяций внутри вида резко увеличивает его приспособительные способности, освоение среды и, в конечном счете, возможности выживания. Все это позволяет считать популяцию элементарной эволюционирующей структурой.

Совокупность популяций различных видов, взаимосвязанных между собой и находящихся в тесном единстве с окружающей средой, составляют экосистему.

При изучении различных экосистем экологи часто исследуют значимость видов в органической среде, их место, отношение к пище и врагам, или, проще говоря – «Элтонскую» нишу. Экологическая ниша – место, занимаемое видом в биоценозе, включающее комплекс его биоценотических связей и требований к факторам среды. Термин введён в 1914 году Дж. Гринеллом и в 1927 году Чарльзом Элтоном. В настоящее время определение Гриннелла принято называть пространственной нишей (по смыслу термин ближе понятию местообитание), а определение Элтона называют трофической нишей (экологическая ниша представляет собой сумму факторов существования данного вида, основным из которых является его место в пищевой цепочке).

Ниши обеспечивают смысловую основу для связи видов, обитающих в одной и той же окружающей среде. Эти виды могут быть расположены вдоль

гипотетической «нишевой оси», которая указывает степень сходства видов друг с другом. Виды с перекрывающимися нишами конкурируют за любые ресурсы, связанные с нишевой осью, и, следовательно, имеют меньшую вероятность гармоничного сосуществования. В тех случаях, когда в качестве нишевой оси можно использовать один ограничивающий ресурс, то такая ситуация станет простой основой для анализа любых экологических сообществ. Однако, во многих случаях виды конкурируют за широкий спектр абиотических и биотических ресурсов, известных не всем. В таких случаях практически невозможно указать ниши для всех видов в сообществе. Но, несмотря на это, можно описать биотическую составляющую ниш вида, используя пищевые сети – сети трофических взаимодействий вида. Эти сети часто описывают антагонистические взаимодействия, такие как хищничество и паразитизм, но могут также включать мутуализмы (опыление, семенной сперматозоид), когда один вид питается другим, обеспечивая при этом репродуктивные процессы. Пищевые сети описывают потоки энергии и биомассы, проходящие через сообщество , показывают экосистемные функции, и могут дать представление об общей ситуации стабильности в сообществе. Таким образом, описание роли видов в пищевых сетях (т.е. как каждый вид участвует в своем сообществе) предоставляет инструментарий для оценки экологических ниш видов, как с точки зрения их требований к выживанию, так и с точки зрения их воздействия на сообщества.

Актуальность темы. Данная работа представляет интерес, поскольку разработанный алгоритм используется для исследования и оценки экологических ниш видов. Такой метод исследования необходимым для выявления значимости вида из взаимосвязей с другими видами и определения наиболее важных видов в экосистеме во избежание их исчезновения. Данный метод является универсальным инструментом исследования значимости элементов различных систем и может применяться для реальных систем данных.

Целью бакалавской работы является анализ данных пищевой цепочки продуцентов, консументов и редуцентов из структуры биосистемы Хвалынского национального парка с помощью методологии PageRank и моделирование полученных результатов в виде графовой модели с целью выявления наиболее значимых элементов структуры.

Для достижения поставленной цели необходимо решить следующие **задачи**:

— Определить основные понятия, необходимые для изучения концепций значимости различных видов в пищевых сетях;
— Изучить концепции значимости различных видов в пищевых сетях по степени, трофическому уровню, мотиву роли и ее центральности;
— Определить основные понятия, необходимые для изучения теоретических основ алгоритма PageRank;
— Изучить оригинальную формулу суммирования для PageRank, матричное представление уравнения суммирования, матрицу Google, вычисление вектора PageRank, степенной метод нахождения собственного значения, субдоминантное собственное значение для матрицы Google, а также проблему PageRank как линейной системы;
— Разработать программу, в основании которой лежит алгоритм ссылочного ссылочного ранжирования PageRank;
— Проанализовать исходные данные с помощью алгоритма PageRank и найти самое «важное» звено в экосистеме Хвалынского национального парка;
— Представить полученные данные графически в виде ориентированного графа;
— Проанализовать полученные результаты и определить наиболее важные виды в экосистеме Хвалынского национального парка;

Структурное содержание бакалаврской работы. Работа состоит из введения, трех разделов, заключения, списка использованных источников и приложения. Общий объем работы составляет более 40 страниц.

Основное содержание работы

Во **введении** представляется актуальность данной темы формулируется цель работы и решаемые задачи, отмечается структурное содержание бакалаврской работы.

В **первом** разделе рассматриваются концепции значимости различных видов в пищевых сетях по *степени, трофическому уровню, роли и ее центральности.*

Роли включают мезомасштабные структуры для описания прямых и

косвенных взаимодействий видов. Некоторые показатели центральности также включают мезомасштабные (т.е. прямые и косвенные взаимодействия) и глобальные сетевые структуры для описания возможности влияния конкретных видов на остальную часть пищевой сети. Эти меры расширяют границы определений (которые учитывают только локальное соседство центральных видов сети), а также рассматривают влияние важных видов через косвенные взаимодействия. Это расширение означает, что прямая связь между степенью и шириной ниши размыта для других мер центральности.

Меры центральности, включающие мезомасштабные сетевые структуры, обычно рассчитываются путем определения набора пищевых цепочек, в которых участвуют фокальные виды, а затем суммирования участия видов в этих цепочках, так же как и на трофическом уровне, усредненном по добыче. Однако, в отличие от трофических уровней, меры центральности также учитывают пищевые цепи, которые не включают фокальные виды, а также виды «выше» фокальных видов, а также те, которые находятся на более низких трофических уровнях. Две таких меры, «центральность взаимности» и «центральность информации», оба количественно определяют частоту, с которой фокальные виды появляются на путях между парами других видов

В то время как центральности взаимности и информации основана на пищевых цепях (мезомасштабных структурах), другие определения центральности основаны на глобальной структуры пищевой сети. Одна из таких мер, центральность собственных векторов, основана на определяющем собственном векторе – собственном векторе, связанном с наибольшим собственным значением – пищевой веб-матрицей. Собственные векторы используются для разложения матриц на ортогональные (полностью некоррелированные) оси – это процесс, лежащий в основе анализа главных компонентов и других методов координации. Определяющий собственный вектор пищевой сети аналогичен первой оси вариации в методе главных компонентов. В такой формулировке центральность видов i является i-й позицией в определении собственного вектора

Во **втором** разделе работы рассматриваются теоретические основы алгоритма PageRank.

PageRank (пэйдж-ранк) – один из алгоритмов ссылочного ранжирова-

ния, разработанных в 1996 году Сергеем Брином и Ларри Пейджем. Алгоритм применяется к коллекции документов, связанных гиперссылками (таких, как веб-страницы из всемирной паутины), и назначает каждому из них некоторое численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов. Вообще говоря, алгоритм может применяться не только к веб-страницам, но и к любому набору объектов, связанных между собой взаимными ссылками, то есть к любому графу.

PageRank – это числовая величина, характеризующая «важность» веб-страницы. Чем больше ссылок на страницу, тем она «важнее». Кроме того, «вес» страницы A определяется весом ссылки, передаваемой страницей B. Таким образом, PageRank – это метод вычисления веса страницы путём подсчёта важности ссылок на неё.

Брин и Пейдж, изобретатели PageRank, начали с простого уравнения суммирования, корни которых фактически взяты из исследований в области библиометрии, анализа структуры цитирования научных работ. PageRank страницы P_i, обозначенный $r(P_i)$, является суммой значений PageRank всех страниц, указывающих на P_i.

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}, \tag{1}$$

где B_{P_i} – это набор страниц, указывающих на P_i и $|P_j|$ – количество внешних ссылок со страницы P_j. Стоит отметить, что PageRank для ссылающихся на страницу $r(P_j)$ вершин в уравнении (1) поделен на число рекомендаций, сделанных P_j'м узлом (обозначается как $|P_j|$). Проблема уравнения (1) состоит в том, что значения $r(P_j)$ PageRank страниц, ссылающихся на страницу P_i, неизвестны. Чтобы обойти эту проблему, Брин и Пейдж использовали итеративную процедуру. То есть, они предполагали, что в начале все страницы имеют одинаковый PageRank (например, $1/n$, где n – количество страниц, содержащихся в индексе Google в мировой сети). На данный момент правило в уравнении (1) используется для вычисления $r(P_i)$ для каждой страницы P_i в индексе. Правило в уравнении (1) применяется последовательно, заменяя значения предыдущей итерации в $r(P_j)$. Мы вводим еще несколько обозначений, чтобы определить эту *итеративную процедуру*. Пусть $r_{k+1}(P_i)$ – это

PageRank страницы P_i на итерации $k+1$. Тогда

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}. \tag{2}$$

Этот процесс начинается с $r_0(P_i) = 1/n$ для всех страниц P_i и повторяется с расчетом, что баллы PageRank в конечном итоге сойдутся к некоторым окончательным стабильным значениям.

Уравнения (1) и (2) вычисляют PageRank по одной странице за раз. Используя *матрицы*, мы заменяем громоздкий символ $\sum$ и далее на каждой итерации вычисляем вектор PageRank, который использует единичный $1 \times n$ вектор для хранения значений PageRank для всех страниц в индексе. Чтобы сделать это, введем матрицу $\mathbf{H}$ размерностью $n \times n$ и вектор строки π^T размерностью $1 \times n$. Матрица $\mathbf{H}$ является строкой нормализованной матрицы *гиперссылок* с $\mathbf{H}_{ij} = 1/|P_i|$, если существует связь от узла i до узла j, и 0 в противном случае. Хотя $\mathbf{H}$ имеет ту же ненулевую структуру, что и бинарная *матрица смежности* для графа, ненулевыми элементами являются вероятности.

Ненулевые элементы строки i, вершины которых ссылаются на i, а ненулевые элементы столбца i, соответствующие вершинам, на которые ссылается i. Введем строковый вектор $\pi^{(k)T}$, который является вектором PageRank на k-й итерации. Используя данную матричную запись, уравнение (2) может быть записано компактно следующим образом:

$$\pi^{(k+1)T} = \pi^{(k)T}\mathbf{H}. \tag{3}$$

Итерации уравнения (3) совпадают с уравнением (2).

Проблему PageRank можно сформулировать двумя способами:

1. Решить следующую проблему *собственных* векторов для π^T.

$$\pi^T = \pi^T\mathbf{G},$$
$$\pi^T\mathbf{e} = 1.$$

2. Решить следующие *линейные однородные системы* для π^T.

$$\pi^T(\mathbf{I} - \mathbf{G}) = \mathbf{0}^T,$$
$$\pi^T\mathbf{e} = 1.$$

В первой системе цель состоит в том, чтобы найти нормализованный *доминирующий левый собственный вектор* $\mathbf{G}$, соответствующий *доминантному собственному значению* $\lambda_1 = 1$. ($\mathbf{G}$ – стохастическая матрица, поэтому $\lambda_1 = 1$.) Во второй системе цель состоит в том, чтобы найти нормализованный левый нулевой вектор из $\mathbf{I} - \mathbf{G}$. Обе системы подчиняются уравнению нормализации $\pi^T\mathbf{e} = 1$, которое гарантирует, что π^T является вектором вероятности. π^T является стационарным вектором Марковской цепи с матрицей переходов $\mathbf{G}$; было проведено много исследований по вычислению стационарного вектора для общей Марковской цепи, которая содержит более десятка методов нахождения π^T. Однако, специфические особенности матрицы PageRank $\mathbf{G}$ делают один численный метод явным фаворитом – степенной метод.

Степенной метод является одним из старейших и простейших итерационных методов нахождения *собственного доминантного значения* и *собственного вектора* матрицы. Поэтому его можно использовать для нахождения стационарного вектора марковской цепи.

В **третьем** разделе описана эмпирическая часть бакалаврской работы – разработка приложения для анализа устойчивости биосистем. Для изучения методологии PageRank были предоставлены данные пищевой цепочки продуцентов, консументов и редуцентов из структуры Хвалынского национального парка. Для наиболее полного понимания материала, введем определения типов в составе лесной экосистемы.

Продуценты – производители продукции, которой потом питаются все остальные организмы, – это наземные зеленые растения, микроскопические морские и пресноводные водоросли, производящие органические вещества из неорганических соединений.

Консументы – это потребители органических веществ. Среди них есть животные, потребляющие только растительную пищу, – травоядные, или пи-

тающиеся только мясом других животных – плотоядные (хищники), а также потребляющие и то и другое – «всеядные».

Редуценты (деструкторы) – восстановители. Они возвращают вещества из отмерших организмов снова в неживую природу, разлагая органику до простых неорганических соединений и элементов. Возвращая в почву или в водную среду биогенные элементы, они, тем самым, завершают биохимический круговорот. Это делают в основном бактерии, большинство других микроорганизмов и грибы. Функционально редуценты – это те же самые консументы, поэтому их часто называют микроконсументами.

Для изучения особенностей внутреннего устройства биотической системы Хвалынского заповедника специалистами биологического факультета СГУ были предоставлены данные пищевых цепочек из 40 видов различных организмов. Они представлены в виде невзвешенного ориентированного графа, где вершинами будут служить виды, а дугами – пары, кто кем питается: реализуется схема "ребро А к В, если А ест В".

Для начала был рассмотрен итеративный метод нахождения PageRank, за основу входных данных взяли пример ориентированного графа с шестью узлами. Код программы написан на языке Python. На вход программы подается матрица Google размерностью $n \times n$, представленная в виде файла в формате csv, где строчки – номера номера вершин, а колонки – вершины, на которые ссылается узел. Каждая ссылка обозначается как $1/n$, где n – общее число ссылок из узла. При отсутствии связи с соответствующей колонкой ставится 0, при связи с одной вершиной – 1. Программа запускается через файл *dxD.py*, в котором предварительно указывается называние подключаемого csv-файла. Для вероятности случайного перехода на другую страницу мы выбрали стандартное значение $D = .85$. Далее при запуске программы данные итерируются T - раз. На выходе получаем вектор PageRang x^T со значениями веса для каждого узла графа.

В дальнейшем было разработано приложение на языке Python на основании алгоритма PageRank для анализа исходных данных и поиска наиболее важных видов животных Хвалынского национального парка, при исчезновении которых произойдет наибольшее нарушение многих пищевых цепочек. На вход данной программы подается пищевая цепь в виде ориентированного

графа с двумя значениями – номером вершины и дугой (парой, кто кем питается), затем запускает алгоритм PageRank, выполняя итерации по каждому узлу и проверяя ребра для направленного графа. Для графической реализации результатов была использована NetworkX – библиотека Python для создания и обслуживания графиков. Для вероятности случайного перехода на другую страницу мы выбрали стандартное значение $D = .85$. Граф итерируется более 10 раз для получения наиболее точных результатов вычисления рангов.

На выходе работы программы мы получаем данные о важности вершин, представленным на рисунке 1, где в левый столбец – это номер вершины, а правый – ее вес. Самые ранжированные состояния - это состояния с высокой плотностью, и, соответственно, высокой степенью важности.

```
40                              :0.0547329051339
26                              :0.0477147321429
38                              :0.0307230803571
21                              :   0.02925
22                              :   0.02925
24                              :0.028453125
23                              : 0.0260625
16                              : 0.0260625
12                              :  0.022875
25                              :0.02213125
13                              : 0.0196875
39                              : 0.0196875
15                              : 0.0196875
27                              :0.0193991071429
35                              :0.0176224553571
30                              :0.0161584821429
33                              :0.01341875
20                              : 0.0133125
11                              : 0.0133125
18                              : 0.0133125
29                              :0.01299375
28                              :  0.011825
37                              :  0.011825
36                              :  0.011825
34                              :0.01086875
31                              : 0.0107625
19                              :  0.010125
9                               : 0.0069375
10                              : 0.0069375
14                              : 0.0069375
17                              : 0.0069375
1                               :   0.00375
0                               :   0.00375
3                               :   0.00375
2                               :   0.00375
5                               :   0.00375
4                               :   0.00375
7                               :   0.00375
6                               :   0.00375
8                               :   0.00375
```

Рисунок 1 – Результат работы алгоритма PageRank для ориентированного графа

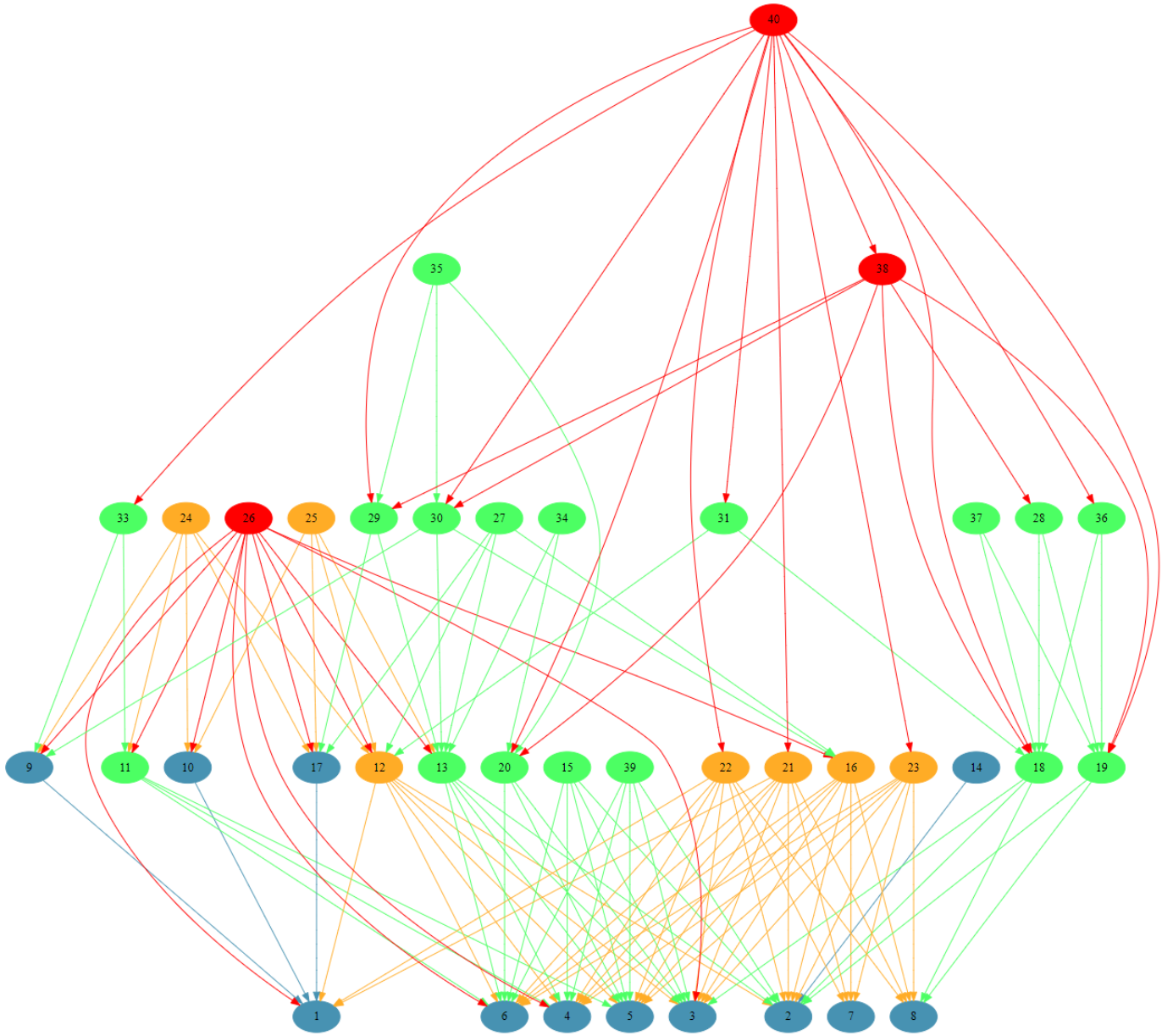

Рисунок 2 – Графическая реализация ориентированного графа PageRank

На основании результата работы программы был визуализирован ориентированный граф, представленный на рисунке 2, и выявлены самые важные виды пищевой цепи, которыми оказались все продуценты (1-8), а также некоторые консументы:

1. 9 – сосновый бражник. Питается сосной;
2. 10 – монашенка. Питается сосной;
3. 14 – жук-олень. Питается дубом;
4. 17 – пилильщик сосновый. Питается сосной.

ЗАКЛЮЧЕНИЕ

В ходе данной работы были определены основные понятия, необходимые для изучения концепций значимости различных видов в пищевых сетях; изучены концепции и основные понятия, необходимые для изучения теоретических основ алгоритма PageRank; определены и изучены теоретические основы алгоритма PageRank, включая оригинальную формулу суммирования для PageRank, матричное представление уравнения суммирования, матрицу Google, вычисление вектора PageRank, степенной метод нахождения собственного значения, субдоминантное собственное значение для матрицы Google, а также проблему PageRank как линейной системы.

В эмпирической части работы был произведен анализ исходных данных с помощью алгоритма PageRank и найдены самые «важные» звенья в экосистеме Хвалынского национального парка, а затем реализованы в виде ориентированного графа. Полученные результаты работы были проанализированы и зафиксированы.