УДК ?????

ИДЕНТИФИКАЦИЯ КЛЮЧЕВЫХ ПРОИЗВЕДЕНИЙ В ЖАНРЕ ФАНТАСТИЧЕСКОЙ ЛИТЕРАТУРЫ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ ГРАФОВ

И.Л. Каширина, В. А. Ковун

*Воронежский государственный университет*

**Аннотация**. Успех книг чаще всего измеряют её тиражом, количеством проданных копий, переизданий, наличием экранизаций и т.п. Такие метрики могут быть недостаточно объективны, так как они подвержены внешним факторам – таким, как маркетинг и актуальность в момент выхода. Более того, эти метрики подвержены искажению с течением времени. Настоящая статья посвящена использованию теории графов для поиска наиболее влиятельных произведений мировой фантастической литературы через построение графа влияния и вычисления различных метрик для такого графа. Также произведён анализ влияния отдельных авторов.

**Ключевые слова.** Теория графов, книги, анализ графов, анализ социальных сетей, центральность, литература

IDENTIFICATION OF MOST INFLUENTAL SCIENCE FICTION LITERATURE WORKS VIA GRAPH ANALYSIS

I.L. Kashirina, V.A. Kovun

Voronezh State University

**Abstract**. A book success is often measured by its printings count, box office, existence of filming, critical acclaim etc. Those measures may lack objectiveness since they may be influenced by external factors, such as trends or advertisement. Moreover, those measures are not capable to capture the impact of a literature work over time. This article is dedicated to usage of graph analysis for identification of the most influential science fiction works via building an influence network and measuring centrality values for literature works represented as a nodes of the graph. The writers’ influence analysis is also performed.

**Keywords.** Optimization methods, neural networks, gradient descent, stochastic gradient, quasi-Newton methods, global and local minimum, objective error function.

**1. Введение.**

Определить самые влиятельные произведения мировой фантастической литературы непросто. Показатели, лежащие на поверхности - количество проданных копий, переизданий, оценки критиков, наличие экранизаций и т.п. - подвержены таким внешним факторам, как маркетинг и актуальность в момент выхода. Для более объективного анализа успеха книги можно использовать метрики, применяемые в анализе социальных сетей. В настоящем исследовании используется измерение успешности произведения через вычисление метрик центральности для построенного из собранных данных графа связей. Для сбора данных использовался популярный русскоязычный сайт с рецензиями на литературные произведения (в основном, из жанра фантастики), для построения графа связей использовались предоставляемые пользователями ссылки на похожие произведения. Используется сочетание четырёх наиболее часто используемых метрик: степень вершины, центральная близость вершины, гармоническая центральность и алгоритм ссылочного ранжирования PageRank.

**2. Материалы и методы.**

*2.1.Сбор данных.*

Для формирования графа влияния использовались данные, собранные с сайта FantLab.ru.

Фантлаб является популярной социальной сетью и онлайн базой данных, содержащей в себе данные о литературных художественных произведениях (преимущественно фантастической направленности). Каждое из произведений в базе сайта содержит различные метаданные – такие, как авторство, год издания, количество и годы переизданий, наличие и авторство переводов на русский язык и так далее. Сайт позволяет пользователям регистрироваться и расширять базу данных, добавляя новую и расширяя имеющуюся информацию о произведениях и оценивая хранимые в базе данных произведения.

Зарегистрированные пользователи также могут добавлять данные по похожести между произведениями и голосовать «за» или «против» предложенных другими пользователями вариантов.

Для предлагаемого анализа используется набор из подможества наиболее популярных произведений всей базы данных сайта Фантлаб. Произведения, не ссылающиеся на другие произведения и не имеющие входящих ссылок, в анализе не участвовали. Для сбора данных был написан парсер сайта, перебирающий идентификаторы книг и использующий XPath для сбора данных со страниц, также использовалось получение JSON с данными по похожим книгам через открытый API сайта с его последующим парсингом. Был собран набор данных по 41838 различных произведений 783 авторов.

По каждой книге были собраны следующие данные: идентификатор, название, автор, год выхода книги и массив идентификаторов похожих произведений.

На рис. 1 приведены наибольшие доли лет выхода произведений в собранном наборе данных.

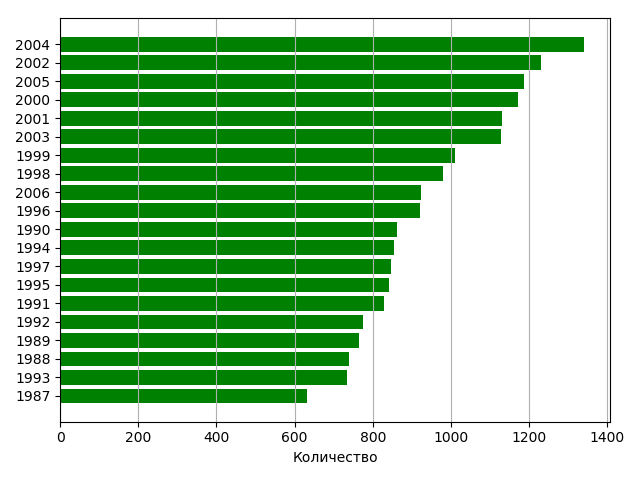


Рис. 1. 20 наиболее часто встречающихся лет выхода произведений в наборе данных.

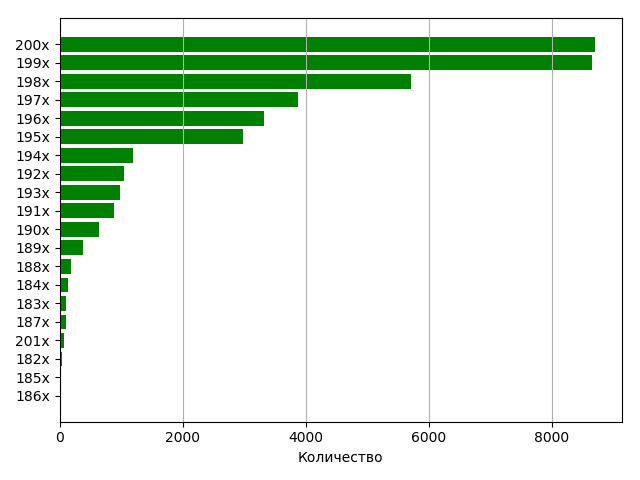


Рис. 2. Доли десятилетий выхода произведений в наборе данных.

*2.2. Анализ собранных данных.*

На сайте, с которого собран набор данных, используется рейтинговая система похожести произведений. Для любого произведения пользователи сайта могут предложить одно или несколько похожих произведений, в то время как остальные пользователи могут голосовать за или против предложенных вариантов. Варианты со слишком низким рейтингом исключаются из выдачи списка похожих произведений. Настоящее исследование исходит из предположения, что указанная пользователями похожесть некоторого произведения на другое вышедшее ранее произведение обозначает, что раннее произведение оказало влияние на более позднее. Однонаправленная связь такого влияния позволяет построить ориентированный граф, вершины которого соответствуют произведениям, и вычислить характеристику центральности вершин для получения оценки исторической важности произведений из набора данных. С другой стороны, нередки ситуации, в которой два или более произведений созданы независимо друг от друга, но при этом под влиянием некоторого третьего. Пользователи могли отметить такие произведения как похожие, что может повлечь за собой ошибочный учёт двух книг с общим источником вдохновения как влияющих одна на другую. Для того, чтобы минимизировать действие таких ситуаций на общую оценку произведений, вычисляется сразу несколько различных метрик центральности, исходя из которых вычисляется итоговая общая оценка.

Пусть – ориентированный граф, состоящий из множества вершин и множества рёбер , представляющих собой упорядоченные пары , обозначающие наличие направленной связи из в . Расстоянием будем называть длину кратчайшего существующего пути из вершины в вершину , при отсутствии такого пути расстояние считается бесконечным. Пусть также – множество всех рёбер, направленных в заданную вершину, а – множество всех рёбер, направленных из заданной вершины. Каждое из произведений в наборе собранных данных соответствует вершине графа , и каждое указание похожего и вышедшего раньше произведения соответствует ребру .

Для определения относительной важности вершин вводится понятие центральности (centrality). Это характеристика, выражающая «влияние» или «важность» той или иной вершины внутри графа в виде вещественного числа. В целом, важность вершины можно трактовать либо как важность типа потока, проходящего через вершину, либо как важность вершины для сохранения связности графа. Однако, поскольку понятия «важности» и «влияния» имеют широкий ряд значений и могут трактоваться по-разному в зависимости от ситуации и решаемой задачи, существует множество различных метрик центральности для графов.

Согласно П. Болди[], меры центральности можно классифицировать по следующим категориям:

- Метрики, основанные на показателе степени вершины;

- Метрики, основанные на числе возможных путей в графе, походящих через заданную вершину;

- Метрики, основанные на кратчайших расстояниях от заданной вершины до остальных вершин;

- Спектральные метрики (spectral indices).

Первые три категории называются геометрическими метриками. Последний класс основан на вычислении собственного вектора для преобразованной матрицы смежности вершин графа.

Для вычисления показателей влияния одних произведений на другие в данной работе используется комбинация следующих четырёх метрик центральности вершины графа.

Степень вершины (центральность по степени) для любой заданной вершины – это количество входящих в вершину рёбер графа:

Центральная близость вершины [1, 2] – величина, показывающая, насколько близко вершина расположена относительно других вершин. Центральная близость определяется как величина, обратная сумме всех конечных расстояний между данной вершиной и всеми остальными вершинами графа:

Гармоническая центральность вершины [1, 2] – величина, равная сумме обратных конечных расстояний от каждой из вершин до заданной:

Метрика PageRank [3], предложенная С. Брином и Л. Пэйджем для использования в поисковой системе Google, полагается на принцип «важности» вершины: вершина тем «важнее», чем больше возможных путей из всех вершин графа в заданную. Кроме того, вес вершины зависит от веса вершин, из которых существуют пути в заданную вершину. Для заданной вершины её PageRank вычисляется следующим образом:

где – коэффициент затухания, обычно принимаемый равным 0.85, и – элемент матрицы смежности графа , то есть:

Каждая из указанных метрик вычисляется отдельно и нормализуется, принимая значение между 0 и 1. Общий рейтинг влиятельности для каждого произведения из набора данных вычисляется как среднее арифметическое всех четырёх вычисленных метрик после нормализации:

Перечисленные метрики выбраны как основные и наиболее часто используемые на практике, а также наиболее хорошо отображающие важность вершины графа как влиятельность соответствующего ей произведения, позволяя отобрать именно те произведения, которые играют наиболее важную роль в истории фантастической литературы.

Центральность по степени отображает количество прямых отсылок, полученных в других произведениях, что является простейшей мерой влияния, используемой, например, в анализе влияния научных статей. Центральная близость и гармоническая центральность, в свою очередь, отображают расстояния от вершины до всех остальных: произведения с высоким показателем этой метрики часто цитируются произведениями, которые, в свою очередь, также часто цитируются. Таким образом, эти оценки дают дополнительную информацию в сравнении с центральностью по степени, обнаруживая произведения, вдохновившие другие влиятельные произведения и, таким образом, сыгравшие важную роль в становлении литературы. PageRank, в свою очередь, подсчитывает взвешенное количество отсылок, полученных произведением, где вес отсылки зависит от количества отсылок, полученных из входящей вершины. Таким образом, более высокий показатель этой метрики получают те произведения, на которые ссылаются лишь важные (в смысле полученных отсылок) вершины.

В целом, каждая из этих метрик отображает различные концепции центральности, и каждая из них важна для обнаружения влиятельных произведений, поэтому на общую оценке вершины в анализе влияют все четыре этих метрики в равной степени.

## 3. Результаты и их обсуждение

## 3.1. Результаты анализа влиятельности произведений.

## Результаты анализа приведены на рис. 1. Можно заметить, что большая часть произведений была выпущена в начале-середине ХХ века, что является частью ожидаемого результата для выбранного метода анализа.

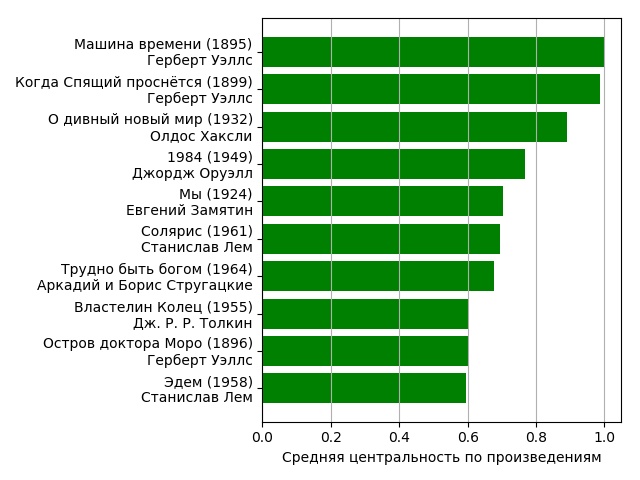
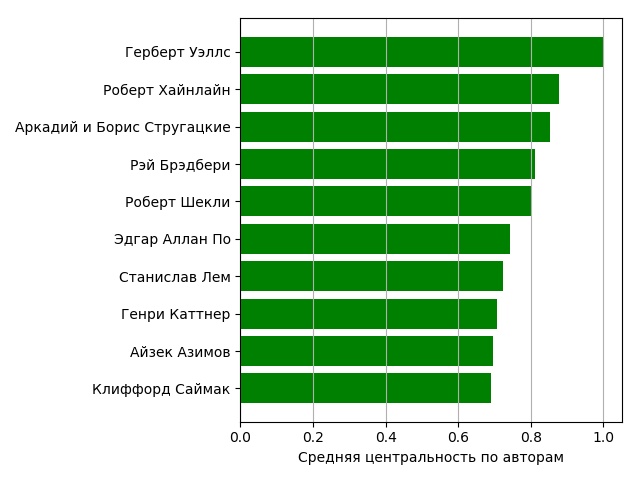


Рис. 1. 10 произведений, соответствующих вершинам графа связей с наивысшей средней центральностью

*3.2 Результаты анализа влиятельности авторов.*

Результаты анализа влиятельности авторов приведены на рис. 2. Анализ данных был произведён при помощи программы, написанной на языке Python 3 с использованием библиоте**к** networkx, matplotlib и numpy.



## Рис. 2. 10 авторов, соответствующих вершинам графа связей с наивысшей средней центральностью

**4. Заключение**

Проведённый анализ взаимосвязи произведений позволяет получить результаты, которые можно считать сравнительно объективными критериями влиятельности произведений на мировую культуру. Данный метод возможно использовать не только на литературных произведениях, но и на любых наборах данных со ссылками – таких, как фильмы, научные статьи и т.п.

**Список литературы**

1. Демченко М.В. Сравнительный анализ и оценка эффективности маркёров атеросклероза магистральных артерий / М.В. Демченко, И.Л. Каширина // Актуальные проблемы прикладной математики, информатики и механики : Сб. тр. Международ. науч.-тех. конференции, Воронеж, 18-20 декабря 2017 г. – Воронеж. : Изд-во «Научно-исследовательские публикации», 2017. – С. 636-643.

2. Jordan J. Intro to optimization in deep learning: Gradient Descent/ J. Jordan // Paperspace. Series: Optimization. – 2018. - URL: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

3. Каширина И.Л. Нейросетевые и гибридные системы: учебно-методическое пособие для вузов / И.Л. Каширина, Т.В. Азарнова. – Воронеж : Издательский дом ВГУ, 2014. – 80 с.

4. Scikit-learn - машинное обучение на Python. – URL:

http://scikit-learn.org/stable/modules/generated/sklearn.neural\_network.MLPClassifier.html

5. Keras documentation: optimizers. – URL: https://keras.io/optimizers

6. Ruder S. An overview of gradient descent optimization algorithms / S. Ruder // Cornell University Library. – 2016. – URL: https://arxiv.org/abs/1609.04747

7. Robbins H. A stochastic approximation method / H. Robbins, S. Monro // The annals of mathematical statistics. – 1951. – Vol. 22. – P. 400-407.

8. Нестеров Ю.Е. Метод минимизации выпуклых функций со скоростью сходимости O(1/k2 ) / Ю.Е. Нестеров // Докл. АН СССР. – 1983. - Т. 269. – № 3. – С. 543-547.

9. Поляк Б.Т. О некоторых способах ускорения сходимости итерационных методов / Б. Т. Поляк // Ж. вычисл. матем. и матем. физ. – 1964. – T. 4. – № 5. – C. 1-17.

10. Kukar M. Cost-Sensitive Learning with Neural Networks / M. Kukar, I. Kononenko // Machine Learning and Data Mining : proceedings of the 13th European Conference on Artificial Intelligence. – 1998. – P. 445–449.

11. Демченко М.В. Построение нейросетевого классификатора для выявления риска атеросклероза магистральных артерий / М.В. Демченко // Оптимизация и моделирование в автоматизированных системах : материалы всероссийской молодежной науч. школы, Воронеж, 13 декабря 2017 г. – Воронеж. : Изд-во Воронежский государств. технический университет, 2017. – С. 29-36.

# 12. Duchi J. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization / J. Duchi, E. Hazan, Y. Singer // The Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2121-2159.

# 13. Zeiler M.D. ADADELTA: An Adaptive Learning Rate Method / Cornell University Library. – 2012. – URL: https://arxiv.org/abs/1212.5701

14. Николенко C. Глубокое обучение / С. Николенко, А. Кадурин, Е. Архангельская. – СПб : Питер, 2018. – 480 с.

# 15. Kingma D.P. Adam: A Method for Stochastic Optimization / D.P. Kingma, J. Ba // Cornell University Library. – 2014. – URL: https://arxiv.org/abs/1412.6980

16. Гудфеллоу Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль. – М. : ДМК Пресс, 2018. – 652 с.

# 17. Поляк Б.Т. Введение в оптимизацию / Б.Т. Поляк. – М. : Наука. Главная редакция физико-математической литературы, 1983. – 384 с.

18. Fletcher R. Practical methods of optimization / R. Fletcher. – Wiley, 2000. – 450 p.

19. Schraudolph N. N. A Stochastic Quasi-Newton Method for Online Convex Optimization / N.N. Schraudolph, J. Yu, S. Gunter // Statistical Machine Learning. – 2017. – URL: http://proceedings.mlr.press/v2/schraudolph07a/schraudolph07a.pdf

20. Ruder S. Optimization for Deep Learning Highlights in 2017 / S. Ruder // Optimization for Deep Learning Highlights in 2017. – 2017. – URL: http://ruder.io/deep-learning-optimization-2017

21. Kawaguchi K. Deep Learning without Poor Local Minima / K. Kawaguchi // Advances in Neural Information Processing Systems. – 2016. – URL: http://arxiv.org/abs/1605.07110

22. Zhang C. Understanding deep learning requires rethinking generalization / C. Zhang, S. Bengio, S. Bengio, M. Hardt, B. Recht, O. Vinyals // Cornell University Library. – 2016. – URL: https://arxiv.org/abs/1611.03530

23. Wilson A.C. The Marginal Value of Adaptive Gradient Methods in Machine Learning / A.C. Wilson, R. Roelofs, M. Stern, N. Srebro, B. Recht // Cornell University Library. – 2017. – URL: https://arxiv.org/abs/1705.08292

24. Тhe MNIST database. – URL: <http://yann.lecun.com/exdb/mnist/>

**СВЕДЕНИЯ ОБ АВТОРАХ**

Каширина Ирина Леонидовна – д. т. н., профессор кафедры математических методов исследования операций факультета ПММ Воронежского государственного университета.

Электронная почта: [kash.irina@mail.ru](mailto:kash.irina@mail.ru)

Демченко Мария Владимировна – аспирант факультета ПММ Воронежского государственного университета

Электронная почта: [masha-vrn@yandex.ru](mailto:masha-vrn@yandex.ru)