УДК 519.173

КОЛИЧЕСТВЕННЫЙ ПОДХОД К ОЦЕНКЕ ВЛИЯНИЯ ЛИТЕРАТУРНЫХ ПРОИЗВЕДЕНИЙ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ ГРАФОВ

В. А. Ковун, И. Л. Каширина

Воронежский государственный университет

Поступила в редакцию 23.08.2019 г.

Аннотация. Успех книг чаще всего измеряют тиражом, количеством проданных экземпляров, переизданий, наличием экранизаций и т. п. Такие метрики могут быть недостаточно объективны, так как они подвержены внешним факторам – таким, как маркетинг и актуальность произведения в момент выхода. Более того, эти метрики подвержены искажению с течением времени. В настоящей статье предлагается количественный подход к оценке значимости литературных произведений через построение графа влияния и вычисления различных показателей центральности для такого графа. Предлагаемый метод применяется к сети, содержащей ссылки между более чем 460000 произведениями мировой фантастической литературы. Из этой сети получен список основных книг, которые можно считать основой жанра. Также приводятся результаты анализа влияния отдельных авторов.
Ключевые слова: теория графов, книги, анализ социальных сетей, метрики центральности, PageRank.

1. ВВЕДЕНИЕ

Во всем мире прилагаются серьезные усилия для выявления наиболее значимых творческих работ в разных областях, включая фильмы, романы, пьесы, стихи, картины и научные исследования. К сожалению, результаты такого анализа часто бывают противоречивы. Проблема заключается в сложности формального определения качества творческой работы.

Наиболее изученными в данный момент являются методы оценки качества научной литературы. Сети, образованные цитатами из научной литературы, давно находятся в центре многих исследований [1, 2]. Исследователи выявили многочисленные метрики, которые определяют, какой документ, автор или журнал является лучшим, наиболее значимым или наиболее влиятельным в своей области. Эти показатели варьируются от простых, таких как общее количество ссылок [3, 4], к сложным, таким как PageRank [5]. Сети научного цитирования предоставляют большие объемы данных для анализа и построения моделей [6].

Как и ученые, писатели часто находятся под влиянием ранее написанных произведений. Однако, в отличие от исследователей, писатели не обязаны ссылаться на эти работы. Выявить наиболее влиятельные произведения мировой литературы непросто. Показатели, лежащие на поверхности – количество проданных экземпляров, переизданий, оценки критиков, наличие экранизаций и т. д. – подвержены воздействию внешних факторов, таких как маркетинг и актуальность на момент выпуска. Если бы можно было получить данные, определяющие ссылки между творческими работами, стало бы возможным применить анализ на основе цитирования для разработки объективной метрики для оценки значимости данной работы. В некоторых творческих областях такие данные сейчас существуют. Например, база данных IMDb содержит крупнейшую цифровую коллекцию метаданных о фильмах и телевизионных программах, от информации о актерах и съемочной группе до критических обзоров. В частности, для каждого фильма

есть раздел под названием «Связи», в котором содержится список ссылок на похожие фильмы. Анализируя эту сеть цитирования, можно исследовать пригодность метрик для оценки значимости фильма на основе распространения влияния в мире кино [8, 12].

Для более объективного анализа успешности книги также можно применять метрики, используемые при анализе социальных сетей [7]. В данном исследовании используется измерение успешности литературного произведения через вычисление метрики центральности для графа связей, построенного из собранных данных. Для сбора данных использовался популярный русскоязычный сайт с обзорами литературных произведений (в основном из жанра фантастической литературы), для построения графа связей использовались ссылки на похожие произведения, предоставленные пользователями сайта. Для анализа значимости применялась комбинация четырех наиболее часто используемых метрик: степень вершины, центральная близость вершины, гармоническая центральность и алгоритм ранжирования ссылок PageRank.

2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Сбор данных

Для формирования графа влияния использовались данные, собранные с сайта FantLab.ru, предоставляющего открытый API для сбора данных.

Фантлаб – популярная социальная сеть и онлайн-база данных, содержащая информацию о литературных художественных произведениях (в основном, фантастической направленности, а также «пограничных» жанрах, таких как магический реализм, готическая проза и пр.). Посещаемость сайта составляет в среднем 791 тыс. посетителей в месяц. На сайте представлено более 462 тысяч произведений и более чем 3500 авторов.

Каждая из книг в базе данных сайта содержит различные метаданные, такие как авторство, год публикации, количество и годы переизданий, наличие и авторство переводов на русский язык и так далее. Сайт позволяет пользователям регистрироваться и расширять базу данных, добавлять новую и дополнять существующую информацию о книгах и оценивать произведения, хранящиеся в базе данных.

Зарегистрированные пользователи также могут добавлять данные о сходстве между произведениями и голосовать «за» или «против» вариантов, предлагаемых другими пользователями.

Для предлагаемого анализа используется набор наиболее популярных произведений и авторов из всей базы данных сайта Фантлаб. Произведения, в которых отсутствуют ссылки на другие книги 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. В результате был сформирован набор данных, содержащий 46838 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 приведен рейтинг, включающий годы, в которые было написано больше всего произведений, представленных в собранном наборе данных, на рис. 2 – рейтинг десятилетий.

2.2. Анализ собранных данных

На сайте, с которого собран набор данных, используется рейтинговая система оценки 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

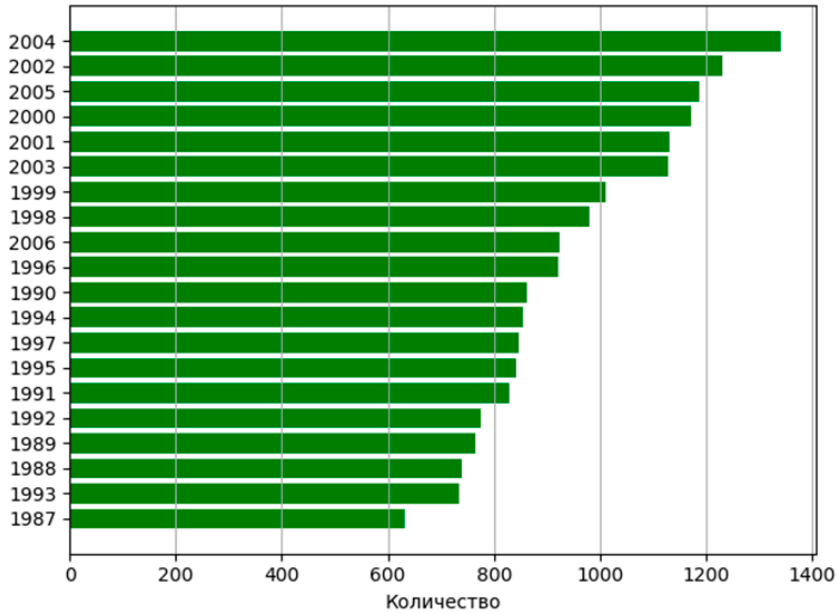

Рис. 1. Двадцать наиболее часто встречающихся лет выхода произведений в наборе данных

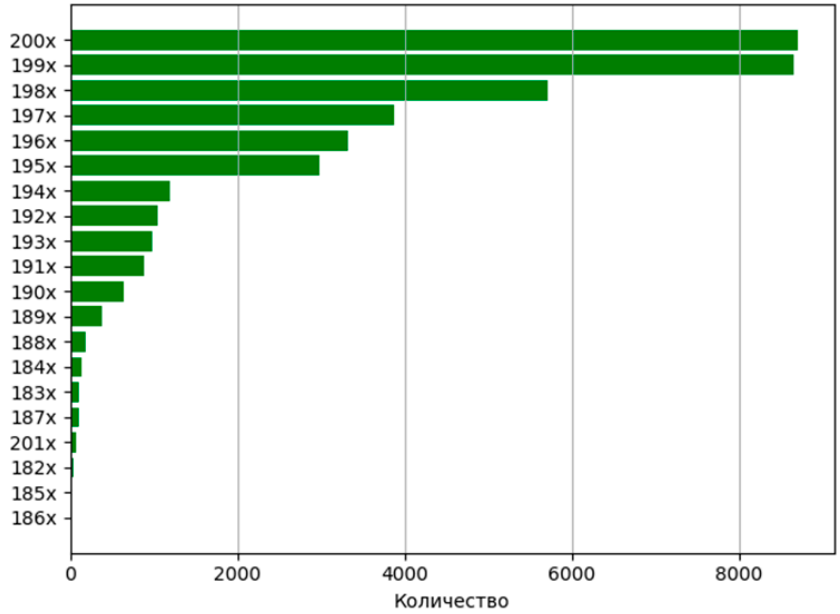

Рис. 2. Рейтинг десятилетий выхода произведений в наборе данных

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 $G = (V, E)$ – ориентированный граф, состоящий из множества вершин V и множества рёбер E, представляющих собой упорядоченные пары $(u, v) : u, v \in V$, обозначающие

наличие направленной связи из u в v. Расстоянием $d(u,v)$ будем называть длину кратчайшего существующего пути из вершины u в вершину v, при отсутствии такого пути расстояние считается бесконечным. Пусть также $N(v) = \{u \in V : (u,v) \in E\}$ – множество всех рёбер, направленных в заданную вершину v, а $O(v) = \{u \in V : (v,u) \in E\}$ – множество всех рёбер, исходящих из заданной вершины v. Каждое из произведений в наборе собранных данных соответствует вершине u графа G, и каждое указание похожего и вышедшего раньше произведения v соответствует ребру (u,v).

На рис. 3 приведён небольшой фрагмент графа влияния, построенного с использованием платформы Cytoscape по 46838 произведениям, представленным на сайте Фантлаб. Дуга, направленная от книги A к книге B означает, что пользователи сайта Фантлаб считают, что книга A похожа на книгу B, при этом книга A вышла позднее книги B.

Аналогичный граф может быть построен не только по произведениям, но и, в целом по авторам. Дуга, направленная от автора A к автору B означает, что у автора A есть хотя бы одна книга, похожая на хотя бы одну книгу автора B, вышедшую ранее.

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 [13], 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.

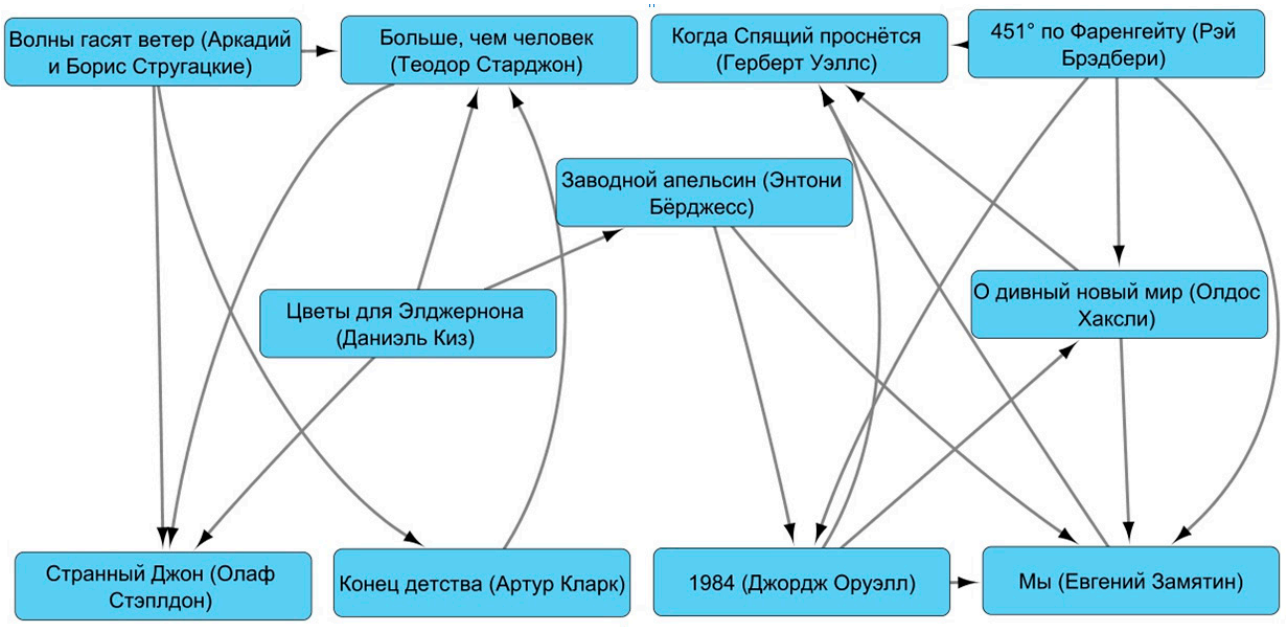

Рис. 3. Фрагмент графа влияния по произведениям

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:

$$I(v) = |N(v)|. \quad (1)$$

Центральная близость вершины [9, 10] – 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:

$$C(v) = \frac{1}{\sum_{u \in V, d(u,v) \neq \infty} d(u,v)}. \quad (2)$$

Гармоническая центральность вершины [10, 13] – величина, равная сумме обратных конечных расстояний от каждой из вершин до заданной:

$$H(v) = \sum_{u \in V, d(u,v) \neq \infty} \frac{1}{d(u,v)}. \quad (3)$$

Метрика PageRank [5, 11], предложенная С. Брином и Л. Пэйджем для использования в поисковой системе Google, относится к центральным метрикам и 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 из следующего соотношения:

$$P(v) = d \sum_{u \neq v} a_{uv} \frac{P(u)}{O(u)} + \frac{1-d}{n}, \quad (4)$$

где $d \in [0,1]$ – коэффициент затухания, обычно принимаемый равным 0.85, и a_{uv} – элемент матрицы смежности графа G, то есть:

$$a_{uv} = \begin{cases} 1, (u,v) \in E \\ 0, (u,v) \notin E \end{cases}. \quad (5)$$

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

$$MeanCentrality(v) = \frac{I(v) + C(v) + H(v) + P(v)}{4}. \quad (6)$$

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 (*in-degree*) отображает количество прямых ссылок, полученных литературным произведением, что является простейшей мерой влияния, используемой, например, в анализе влияния научных статей. Центральная близость (*closeness*) и гармоническая центральность (*harmonic*), в свою очередь, количественно по-разному определяют расстояния от данной вершины до всех остальных: произведения с высоким показателем этих метрик 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 ссылок, полученных произведением, где вес каждой ссылки зависит от количества ссылок, полученных исходящей вершиной. Таким образом, более высокий показатель этой метрики получают те произведения, на которые ссылаются лишь важные (в смысле полученных ссылок) вершины.

В целом, эти метрики отображает различные концепции центральности, и каждая из них важна для обнаружения влиятельных произведений, поэтому в данном исследовании в качестве базовой метрики используется среднее арифметическое представленных четырех метрик. На рис. 4 приведены матрицы корреляций Спирмена между различными видами метрик центральности среди 200 про-

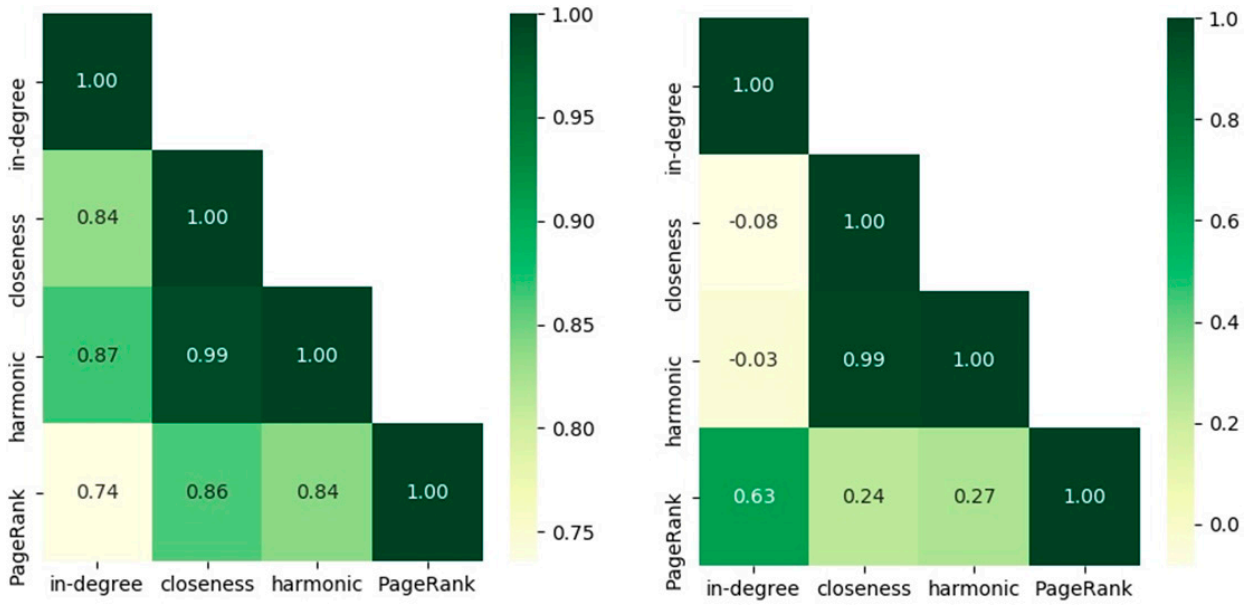

Рис. 4. Матрицы корреляций Спирмена между различными метриками центральности

изведений с наибольшей средней центральностью (справа) и 100 авторами с наибольшей средней центральностью (слева).

Корреляция Спирмена определяет степень тесноты линейной связи между ранжированиями, получаемыми с помощью различных метрик центральности. По рис. 4. можно заметить, что центральная близость и гармоническая центральность приводят к очень похожим спискам лучших книг и авторов, но между ними все же существуют некоторые различия. Также можно отметить, что *PageRank* показывает очень низкую связь с центральной близостью и гармонической центральностью для рангов произведений, а у центральности по степени связь с этими метриками практически отсутствует (из-за ее смещения в более современные произведения), что еще раз подчеркивает важность одновременного учета всех метрик для получения более объективной итоговой оценки.

3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

3.1. Результаты анализа влиятельности произведений

Результаты анализа, проведенного с помощью предлагаемой выше методики, приведены в табл. 1. Анализ данных был произведён при помощи программы, написанной на языке Python 3 с использованием библиотек networkx, matplotlib и numpy.

Можно заметить, что большая часть наиболее влиятельных произведений была выпущена в начале-середине XX века, что является частью ожидаемого результата для выбранного метода анализа. Кроме того, заслуживает внимания факт, что на второй строчке рейтинга расположился не самый известный роман Герберта Уэллса. Однако эксперты отмечают ключевое влияние этой книги на дальнейшее развитие всего жанра фантастики. Одна из причин заключается в том, что она является, по сути, первым романом-антиутопией, предвестником известнейших произведений, расположившихся на третьей, четвертой и пятой строчках полученного рейтинга.

3.2 Результаты анализа влиятельности авторов

Результаты анализа влиятельности авторов приведены в табл. 2. В рейтинге авторов также отсутствуют современные писатели-фантасты, по причине того, что они еще не успели обзавестись большим числом «подражателей» и повлиять на создание других значимых произведений. На первом месте предсказуемо расположился Герберт Уэллс, как основоположник жанра фантастической литературы. Роберт Хайнлайн, оказавшийся на

Таблица 1

Десять произведений, соответствующих вершинам графа связей с наивысшей средней центральностью

Ранг	Произведение	Нормализованная средняя центральность
1	«Машина времени» Герберт Уэллс (1895)	1.000
2	«Когда Спящий проснётся» Герберт Уэллс (1899)	0.887
3	«О дивный новый мир» Олдос Хаксли (1932)	0.875
4	«1984» Джордж Оруэлл (1949)	0.706
5	«Мы» Евгений Замятин (1924)	0.669
6	«Солярис» Станислав Лем (1961)	0.659
7	«Остров доктора Моро» Герберт Уэллс (1896)	0.627
8	«Трудно быть богом» Аркадий и Борис Стругацкие (1964)	0.616
9	«Властелин Колец» Дж. Р. Р. Толкин (1955)	0.605
10	«Война миров» Герберт Уэллс (1898)	0.602

Таблица 2

Десять авторов, соответствующих вершинам графа связей с наивысшей средней центральностью

Ранг	Автор	Нормализованная средняя центральность
1	Герберт Уэллс	1.0
2	Роберт Хайнлайн	0.878
3	Аркадий и Борис Стругацкие	0.852
4	Рэй Брэдбери	0.811
5	Роберт Шекли	0.801
6	Эдгар Аллан По	0.743
7	Станислав Лем	0.723
8	Генри Каттнер	0.708
9	Айзек Азимов	0.696
10	Клиффорд Саймак	0.690

втором месте, на сегодняшний день является самым титулованным писателем-фантастом, получившим шесть премий Хьюго. При этом, если брать в расчет только прямые ссылки (центральность по степени), то среди авторов лидируют Аркадий и Борис Стругацкие, что ожидаемо для русскоязычного сайта.

4. ЗАКЛЮЧЕНИЕ

Проведённый анализ позволяет получить результаты, которые можно считать сравнительно объективными критериями влиятельности произведений и авторов на мировую литературу в жанре фантастики и связанных жанрах. В отличие от многих других исследований влияния конкретных авторов или произведений, предлагаемый анализ опирается на обширные данные, собранные тысячами пользователей, а не только на личный опыт экспертов, и, таким образом, охватывает значительно больший объем книг, чем представлено в других источниках. Однако использование таких данных также вводит определенные ограничения: будучи веб-сайтом на

русском языке, Фантлаб отображает предпочтения только русскоязычных пользователей. Кроме того, ссылки на похожие книги более неформальны, чем традиционные контексты цитирования (например, научные цитаты официально документированы в самих публикациях). Этот неформальный характер границ, многомерность типов отношений и быстрое развитие сети расширяют актуальность представленного метода даже за пределами области художественной литературы.

СПИСОК ЛИТЕРАТУРЫ

1. *Price, D. J.* Networks of scientific papers / D. J. Price // Science. – 1965. – № 149 (3683). – P. 510–515.

2. *Redner, S.* Citation statistics from 110 years of Physical Review / S. Redner // Phys Today. – 2005. – № 58(6). – P. 49–54.

3. *Radicchi, F.* Universality of citation distributions: Toward an objective measure of scientific impact / F. Radicchi, S. Fortunato, C. Castellano // Proc Natl Acad Sci USA. – 2008. – № 105 (45). – P. 17268–17272.

4. *Garfield, E.* The history and meaning of the journal impact factor / E. Garfield // JAMA. – 2006. – № 295(1). – P. 90–93.

5. *Chen, P.* Finding scientific gems with Google's PageRank algorithm / P. Chen, H. Xie, S. Maslov, S. Redner // J Informetrics. – 2007. – № 1(1) – P. 8–15.

6. *Каширина, И. Л.* Модели и численные методы оптимизации формирования эффективной сетевой системы с кластерной структурой / И. Л. Каширина, Я. Е. Львович, С. О. Сорокин // Информационные технологии. – 2015. – Т. 21, № 9. – С. 657–662.

7. *Лесковец, Ю.* Анализ больших наборов данных / Лесковец Ю., Раджараман А., Джеффри Д. Ульман. – М. : ДМК Пресс, 2016. – 498 с.

8. *Bioglio, L.* Identification of key films and personalities in the history of cinema from a Western perspective / L. Bioglio, R. G. Pensa // Applied Network Science – Dijon, 2018. – Электрон. журн. – Режим доступа: https://doi.org/10.1007/s41109-018-0105-0

9. *Freeman, L. C.* A set of measures of centrality based upon betweenness / L. C. Freeman // Sociometry. – 1977. – № 40. – С. 35–41.

10. *Щербакова, Н.* Меры центральности в сетях. / Н. Щербакова // Проблемы информатики. – Новосибирск, 2015. – № 2. – С. 18–30.

11. *Brin, S.* The anatomy of a large-scale hypertextual web search engine. / S. Brin, L. Page // Computer Networks and ISDN Systems. – Amsterdam, 1998. – №30(1). – P. 107–117.

12. *Canet, F.* Quantitative approaches for evaluating the influence of films using the imdb database / F. Canet, M. Á. Valero, L. Codina// Communication & Society, 2016. – № 29(2). – P. 151.

13. *Boldi, P.* Axioms for centrality / P. Boldi, S. Vigna // Internet Math. – 2014. – № 10(3-4) – P. 222–262.

14. *Newman, M.* The structure and function of complex networks / M. Newman // SIAM Rev Soc Ind Appl Math. – 2003. – 45(2) – P. 167–256.

15. *Chen, C.* Tracing knowledge diffusion / C. Chen, D. Hicks// Scientometrics. – 2004. – № 59(2) – P. 199–211.

16. *Ковун, В. А.* Идентификация ключевых произведений в жанре фантастической литературы с использованием теории графов / В. А. Ковун, И. Л. Каширина // В сборнике: Информатика: проблемы, методология, технологии Сборник материалов XIX международной научно-методической конференции. Воронеж, 2019. – С. 1474–1478.

Каширина Ирина Леонидовна – д-р техн. наук, профессор кафедры математических методов исследования операций факультета ПММ Воронежского государственного университета, e-mail: kash.irina@mail.ru

Ковун Владислав Анатольевич – аспирант факультета ПММ Воронежского государственного университета, e-mail: sidav94@gmail.com

QUANTITATIVE APPROACH TO EVALUATING THE INFLUENCE OF LITERARY WORKS USING THE GRAPH THEORY

V. A. Kovun, I. L. Kashirina

Voronezh State University

Annotation. The success of books is most often measured by circulation, the number of copies sold, reprints, the availability of adaptations, etc. Such metrics may not be objective enough, since they are subject to external factors - such as marketing and the relevance of the work at the time of release. Moreover, these metrics are subject to distortion over time. This article proposes a quantitative approach to assessing the significance of literary works through the construction of a graph of influence and the calculation of various centrality indicators for such a graph. The proposed method is applied to a network containing links between more than 460,000 works of world science fiction. A list of the main books that can be considered the basis of the genre is obtained from this network. The results of the analysis of the influence of individual authors are also presented.
Keywords: graph theory, books, analysis of social networks, centrality metrics, PageRank.

Kashirina I. L. – Doctor of Technical Sciences, Professor, Department of Mathematical Methods Operations Research, Faculty of Applied mathematics, Informatics and mechanics, Voronezh State University, e-mail: kash.irina@mail.ru

Kovun V. A. – postgraduate student of the Department of Mathematical Methods Operations Research, Faculty of Applied mathematics, Informatics and mechanics, Voronezh State University, e-mail: sidav94@gmail.com