Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2018, №1, Том 10 / 2018, No 1, Vol 10 https://esj.today/issue-1-2018.html

URL статьи: https://esj.today/PDF/08ITVN118.pdf

Статья поступила в редакцию 26.01.2018; опубликована 16.03.2018

Ссылка для цитирования этой статьи:
Краснов Ф.В., Макаров И.А. Прогнозирование развития соавторства в написании научных статей научно-технического центра Газпромнефть на основе модели // Вестник Евразийской науки, 2018 №1, https://esj.today/PDF/08ITVN118.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

For citation:
Krasnov F.V., Makarov I.A. (2018). Predicting co-author relationship for Science and Technology Center of Gazpromneft based on the graph modeling. *The Eurasian Scientific Journal*, [online] 1(10). Available at: https://esj.today/PDF/08ITVN118.pdf (in Russian)

Исследование И.А. Макарова выполнено за счет гранта Российского научного фонда (проект № 17-11-01294) в Национальном исследовательском университете "Высшая школа экономики", Москва

УДК 316.452

Краснов Федор Владимирович
ООО «Газпромнефть НТЦ», Санкт-Петербург, Россия
Эксперт
Кандидат технических наук
E-mail: Krasnov.FV@Gazprom-Neft.ru
ORCID: http://orcid.org/0000-0002-9881-7371
РИНЦ: https://elibrary.ru/author_profile.asp?id=855886

Макаров Илья Андреевич
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
Международная лаборатория прикладного сетевого анализа
Старший преподаватель
Младший научный сотрудник
E-mail: iamakarov@hse.ru
ORCID: http://orcid.org/0000-0002-3308-8825
РИНЦ: https://elibrary.ru/author_profile.asp?id=826008
Researcher ID: http://www.researcherid.com/rid/G-9195-2015
SCOPUS: http://www.scopus.com/authid/detail.url?authorId=56233657100

Прогнозирование развития соавторства в написании научных статей научно-технического центра Газпромнефть на основе модели

Аннотация. Коллективное соавторство в написании научных статей имеет детерминированную и случайную структурные составляющие. Кроме рациональных аспектов при образовании коллектива соавторов отдельной научной статьи существуют и эмоциональные составляющие. Во временной перспективе складываются и распадаются рабочие группы исследователей, обновляется трудовой коллектив и состав подрядчиков, которые участвуют в совместных отраслевых коллаборациях для проведения исследований.

Несмотря на всю сложность соавторства, существуют несколько классов моделей для симуляции образования соавторства. В их числе модели на основании случайных графов и модели образования соавторств на основе компетенций соавторов. Оба математических аппарата разработаны и применяются в течении нескольких десятков лет по отдельности. Но практических применений моделей соавторств в корпоративной практике не так много.

Авторы выдвинули гипотезу о том, что необходимо объединить несколько различных типов моделей для того, чтобы лучше понять природу научных коллабораций в отдельной организации.

Авторы данного исследования поставили задачу разработать методику построения модели соавторства для научно-технического центра, учитывающую различные структурные составляющие соавторства.

В результате авторы разработали модель с использованием методов машинного обучения, случайных графов и модели компетенций. На основании разработанной модели сделан прогноз развития соавторства в написании научных статей научно-технического центра Газпромнефть.

Практическая ценность результатов данного исследования состоит в следующем:

1. Количественно оценен вклад различных структурных составляющих в формировании соавторств при написании научных статей.
2. Прогнозирование развития соавторства в написании научных статей позволяет осуществить планирование корпоративных ресурсов для поддержания роста научных публикаций.

Понимание кластерной структуры соавторства позволяет производить выравнивание направлений научной деятельности в соответствии со стратегическим планом развития научно-технического центра.

Ключевые слова: граф соавторства; модель организационного процесса; метрики графа; методы машинного обучения; прогнозирование узлов графа

Введение

Измерение деятельности научно-исследовательских организаций на основании графа соавторства является хорошо зарекомендовавшей себя практикой [1, 2, 3, 4]. Например, в работе [2] исследователи показывают возможности выявления наиболее производительных авторов («Highly Productive Authors») и влиятельных авторов («Influential Authors»). В работе [4] внимание сосредоточено на выявлении рабочих групп на основе факторного анализа графа соавторства. Исследования графов соавторства проводятся как на микроуровне [5] (организация), так и на макроуровне [6] (глобальные научные сообщества).

Почти в каждой работе, использующей граф соавторства, присутствует изображение графа (рисунок 1).

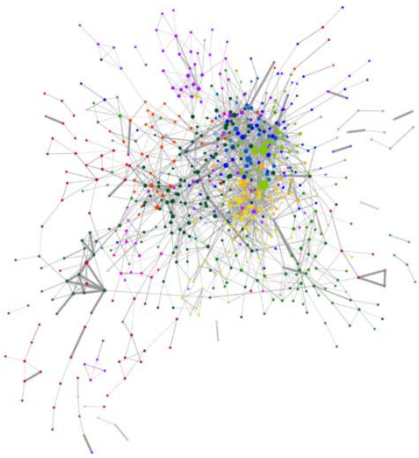

Рисунок 1. *Граф соавторства из работы [7]*

Для моделирования графов соавторства как социальной сети широко применяется следующие стохастические подходы:

1. Случайные графы [8].
2. Модель «Маленький мир» (small-world) [9].
3. Модели, основанные на концепции преимущественного присоединения (preferential attachment) [10, 11].

Одним из существенных ограничений стохастических моделей является фиксированное количество рассматриваемых вершин, или их постоянный рост. На практике в организации изменяется количество потенциальных соавторов.

Так же важно понимать, что стохастические модели ставят целью моделировать граф с определенными параметрами. Такими как, кластеризация, плотность и др.

С другой стороны, формирование малых групп, к которым относится и группа соавторов научной статьи, моделируется на основании принципа дополнительных компетенций [12], относящегося к классу детерминированных методов создания графов соавторства.

Задача предсказания новых вершин графа соавторства рассмотрена в работе [13], а в исследовании [14] изучено предсказание временных свойств новых ребер для графа соавторства большого научно-исследовательского университета. В последней статье применяется комбинированный подход на основе машинного обучения с предварительным отбором признаков авторов, статистических показателей активности за несколько последних временных промежутков, а также структурные индексы влияния и локальные метрики в сети соавторства. Полученные в статье результаты позволили сделать вывод о применимости методов предсказания связей для анализа коллаборативных шаблонов поведения в крупной организации с динамической структурой коллектива, а также меняющимися внешними и внутренними факторами, влияющими на индивидуальную и коллективную публикационную активность.

Сформулируем исследовательский вопрос данной работы в широком смысле:

И.В.1. Какова точность прогноза изменения графа соавторства для одной организации в заданных условия?

Данное исследование состоит из описания методики, сбора и подготовки данных для проведения цифрового эксперимента, результатов эксперимента и заключения.

Методика

Основой для прогнозирования изменений графа соавторств для научно-технического центра являются следующие компоненты:

1. Текущая структура графа соавторств.
2. Внешние по отношению к научно-техническому центру воздействия.
3. Внутренние изменения научно-технического центра.

Рассмотрим каждую из компонент подробнее.

Текущая структура графа соавторств представляет набор метрик, описывающих данный граф соавторств. К таким метрикам согласно работе [15] принято относить следующие:

1. Для ребер:

a. Common Neighbours (CN)
b. Salton Index (SI)
c. Jaccard Index (JI)
d. Hub Promoted Index (HPI)
e. Hub Depressed Index (HDI)
f. Leicht-Holme-Newman Index (LHN1)
g. Preferential Attachment Index (PA)
h. Adamic-Adar Index (AA)
i. Resource Allocation Index (RA)

2. Для вершин:
   a. Degree centrality
   b. Betweenness centrality
   c. Closeness centrality
   d. Harmonic centrality
   e. Clustering

Каждая из этих метрик представляет характерный набор признаков (features) графа соавторств, влияющих на прогноз его изменений.

Внешние по отношению к научно-техническому центру воздействия заключаются в публикационной политике редакций, публикующих научные статьи. В простейшем случае отсутствие возможности опубликовать статью из-за ограничений по объему выпуска журнала приводит несостоявшемуся соавторству. Основные зависимости публикационной активности научно-технического центра от редакций рассмотрены в работе [16]. В данной работе мы ограничились рассмотрением двух ежегодных конференций и одного журнала с двенадцатью выпусками в год.

Внутренние изменения научно-технического центра вызваны изменениями в составе персонала. В организацию приходят новые сотрудники, некоторые сотрудники увольняются. В процессе наставничества и обучения сотрудники приобретают новые компетенции. В результате проведения НИР рождаются новые исследования и научные заделы. Часто, изменения внутренних требований к качеству публикаций также могут стать причиной структурных изменений, подтверждая принцип “publish or perish”, и влияя как на структуру коллектива, так и на параметры активности отдельных сотрудников и исследовательских коллективов.

Рассмотрим подробнее в чем состоит прогноз развития графа соавторств для научно-технического центра. Под развитием мы будем понимать возникновение новых вершин и ребер. Граф соавторств может рассматриваться как накопленным итогом за период, так и инкрементальными изменениями по годам.

Далее будем рассматривать факт авторства, как признак вершины графа соавторств. Другими словами, сотрудник, представляемый вершиной графа соавторства, может как написать, так и не написать статью в следующем временном периоде. Процесс прогнозирования в данном случае будет решать задачу бинарной классификации. Для каждого сотрудника будет определяться вероятность создания статьи по определенной тематике.

Статья является коллективным усилием работы соавторов, обладающих определенным набором компетенций, нашедших свое применение в цели исследования. В этом состоит основная идея принципа дополнительности компетенций. Авторы с одинаковыми компетенциями не имеют рационального обоснования для объединения с целью проведения научного исследования. Будем считать, компетенции атрибутами вершин графа. Для выявления компетенций необходимых для написания статьи будем использовать ключевые слова, а при их отсутствии – метод тематического моделирования текста работы [17].

Результаты эксперимента

В качестве объекта исследования была выбрана публикационная активность НТЦ «Газпромнефть». Данные были получены из открытой электронной библиотеки OnePetro международного сообщества нефтегазовых инженеров (SPE). После очистки было получено 172 статьи. Распределение авторов по годам отображено на рисунке (рисунок 2).

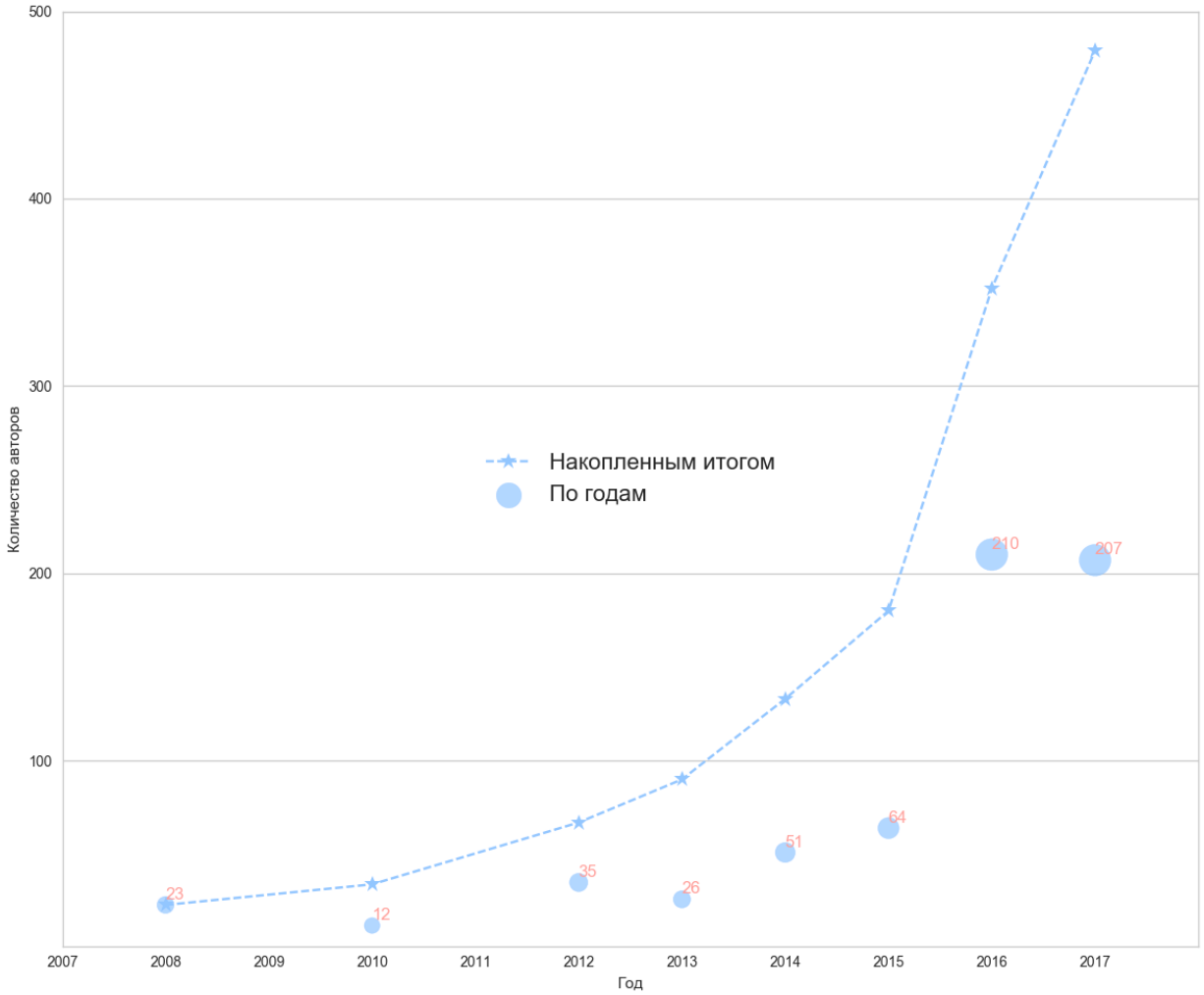

Рисунок 2. *Распределение авторов по годам (разработано авторами)*

Прямолинейным ответом на поставленный исследовательский вопрос может быть интерполяция кривой роста количества авторов. В результате такой оценки получим следующую зависимость $y = 13.3816e^{0.3422x}$ с достоверностью $R^2 = 0.98$ дающей прогноз 585 авторов в 2018 году. Но из графика мы так же видим, что количество авторов в 2017 году (207) меньше, чем в 2016 году (210), что может оказаться насыщением роста и повлиять на прогноз.

Построим прогноз на основании графа соавторства. Для этого построим двудольный граф соавторства с вершинами: автор (479) и статья (171). Авторы обладают техническими компетенциями, статьи характеризуются названием, годом издания и ключевыми словами.

Полученный граф соавторства имеет 26 связанных компонент наибольшая из которых содержит 556 вершин, а остальные – не более 8. Малые связанные компоненты относятся к авторам, написавшим свою первую статью. Наличие малых связанных компонент можно рассматривать, как одну из составляющих роста графа соавторств. В таблице (таблица 1) приведены количества и размеры связанных компонент за каждый год нарастающим итогом.

Таблица 1

Размеры связанных компонент графа соавторств по годам нарастающим итогом

Год	Размеры связанных компонент	Доля малых компонент
2017	**556**, 8, 8, 8, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2	15 %
2016	**367**, 8, 8, 8, 8, 8, 6, 5, 5, 5, 5, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2	23 %
2015	**89**, 22, 21, 15, 12, 12, 8, 8, 8, 8, 6, 6, 5, 4, 3, 3, 2, 2, 2, 2, 2	63 %
2014	**46**, 18, 15, 12, 12, 10, 8, 8, 8, 7, 6, 5, 4, 4, 3, 2, 2, 2, 2, 2	74 %
2013	**23**, 15, 12, 11, 10, 8, 8, 7, 5, 4, 4, 4, 2, 2, 2	80 %
2012	**15**, 14, 12, 11, 8, 8, 7, 4, 4, 4	83 %
2010	**12**, 9, 8, 8, 4, 3	73 %
2008	**12**, 8, 7, 3	60 %

Разработано авторами

Таким образом мы видим, что граф соавторства прогрессирует в сегменте малых связанных компонент по количеству и вместе с тем граф становится более связанным – увеличивается количество узлов в главной связанной компоненте. Для уточнения прогнозирования целесообразно будет учесть такое строение.

На (рисунок 3) приведена инкрементальная динамика прироста графа соавторства по годам. Уточним, что граф соавторств 2017 года является суммой всех изображенных на (рисунок 3).

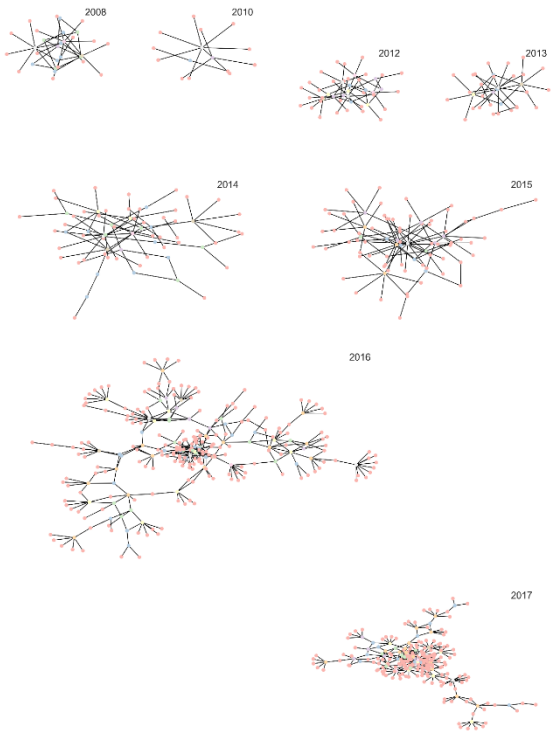

Рисунок 3. *Динамика прироста развития графа соавторств по годам (разработано авторами)*

08ITVN118
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Из рисунка (рисунок 3) можно сделать качественной вывод об увеличении ежегодно прибавляемых к графу соавторства связей. Изменение роста ведет к усложнению структуры графа соавторства в 2016 году, что можно констатировать как «эффект локтя» [18].

Для прогнозирования авторства будем использовать следующие метрики вершин графа:

- Degree centrality;
- Betweenness centrality;
- Closeness centrality;
- Harmonic centrality;
- Clustering.

Распределения метрик вершин графа соавторства приведены на рисунке (рисунок 4).

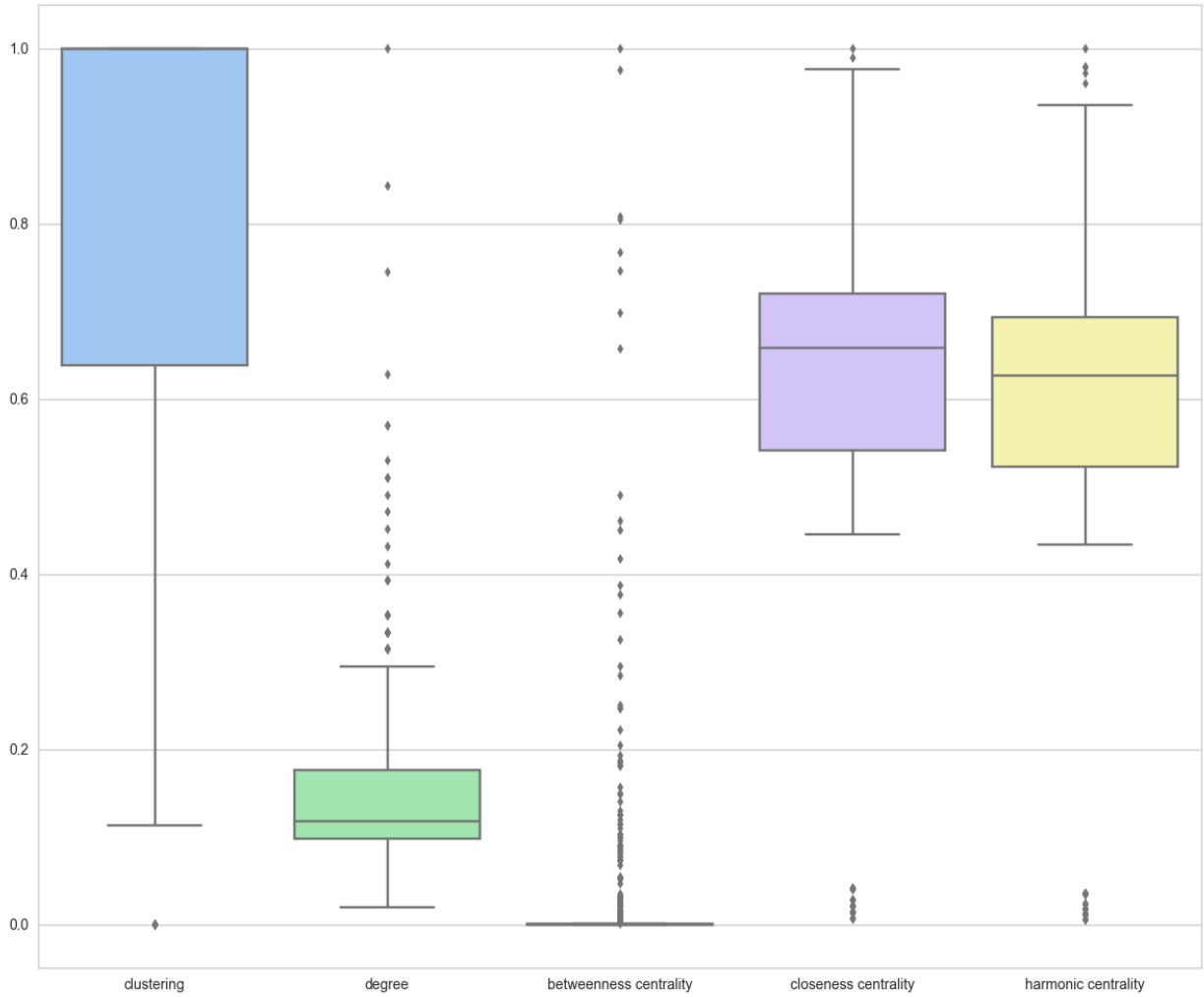

Рисунок 4. *Метрики вершин графа соавторства (разработано авторами)*

Для прогнозирования авторства будем использовать модель бинарной классификации. Выбор модели будем производить на основе ROC-кривой. Обучения модели будем производить на метриках 2016 года. Параметры моделей оптимизированы с помощью кросс-валидации с 5-кратным фолдингом. В результате сравнения различных классификаторов были получены следующие результаты (таблица 2).

Таблица 2

Сравнение классификаторов по метрике ROC AUC

Модель	ROC AUC
KNeighborsClassifier	**0.66**
RidgeClassifier	**0.73**
RandomForestClassifier	**0.72**
SVM	**0.70**
Multi-layer perceptron	**0.75**

Разработано авторами

Лучшее значение метрики ROC AUC показал классификатор на основе нейронной сети (Multi-layer perceptron) с одним слоем из 10 персептронов.

Отчет о выполнении прогноза авторства в 2018 году на основе графа соавторства 2017 года приведен в таблице (таблица 3).

Таблица 3

Отчет о выполнении прогноза авторства на 2018 г.

	precision	recall	f1-score	support
not author	0.80	0.98	0.88	66
author	0.80	0.20	0.32	20
Avg/Total	0.80	0.80	0.75	86

Разработано авторами

В результате прогнозирования определено 406 авторов в 2018 году. Если добавить к этому прогнозу сотрудников, которые напишут свою первую статью в 2018 году, то на основании оценки по динамике роста связанных компонент получим прибавку в 15 %. Итого в 2018 году авторами станут 467 сотрудников.

Заключение

В настоящей работе авторами проведен анализ динамики графа соавторства для одной организации на основании публичных данных о публикациях. Основным аналитическим инструментом авторами выбран двудольный граф соавторства.

В работе применен много компонентный подход к прогнозированию изменению свойств графа соавторства. Анализ малых связанных компонент позволил выявить их долю в ежегодном увеличении количества авторов. Отметим, что доля малых компонент в рассматриваемом графе соавторства уменьшается со временем, что является структурным ограничением роста рассматриваемой организации.

В 2016 году обнаружен «Эффект локтя» – резкое усложнение характера роста графа соавторства по годам.

Авторами сделан прямой прогноз роста на основании тренда роста авторов по годам и уточненный прогноз роста графа соавторства на основе моделирования с помощью методов машинного обучения.

Проведенное сравнение точности классификаторов определило классификатор на основе нейронной сети как наиболее точный для данной задачи.

Прогноз, сделанный на основе модели, показал результат (467) существенно меньший чем результат на основе тренда (585).

В результате проведенного исследования авторы сделали вывод о наличии в структуре графа соавторств важной информации о развитии графа соавторств, которая определяет прогноз роста. Дальнейшее исследование позволит определить значимые признаки образования новых коллабораций, а также регрессионного предсказания новых связей между уже сформировавшимися исследовательскими коллективами. Использование методов векторизации графовых моделей [19] в комбинации с грамотным извлечением признаков позволит улучшить точность предсказания появления новых связей, а также качественно измерить публикационную активность на основе публично доступных метрик журналов и конференций [20].

ЛИТЕРАТУРА

1. Abbasi A., Chung K.S.K., Hossain L. Egocentric analysis of co-authorship network structure, position and performance // Information Processing & Management. – 2012. – Т. 48. – № 4. – С. 671-679.
2. Börner K. et al. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams // Complexity. – 2005. – Т. 10. – № 4. – С. 57-67.
3. Li E.Y., Liao C.H., Yen H.R. Co-authorship networks and research impact: A social capital perspective // Research Policy. – 2013. – Т. 42. – № 9. – С. 1515-1530.
4. Perianes-Rodríguez A., Olmeda-Gómez C., Moya-Anegón F. Detecting, identifying and visualizing research groups in co-authorship networks // Scientometrics. – 2010. – Т. 82. – № 2. – С. 307-319.
5. Chang H.J., Wang W.M. The Hidden Power of Social-Linkage in the Office: A Co-authorship Network Analysis // Proceedings of the 4th Multidisciplinary International Social Networks Conference on ZZZ. – ACM, 2017. – С. 4.
6. Munoz D.A., Queupil J.P., Fraser P. Assessing collaboration networks in educational research: A co-authorship-based social network analysis approach // International Journal of Educational Management. – 2016. – Т. 30. – №. 3. – С. 416-436.
7. Claudel M. et al. An exploration of collaborative scientific production at MIT through spatial organization and institutional affiliation // PloS one. – 2017. – Т. 12. – №. 6. – С. e0179334.
8. Erdos P., Rényi A. On the evolution of random graphs // Publ. Math. Inst. Hung. Acad. Sci. – 1960. – Т. 5. – № 1. – С. 17-60.
9. Watts D.J., Strogatz S.H. Collective dynamics of ‘small-world’networks // nature. – 1998. – Т. 393. – №. 6684. – С. 440-442.
10. Barabási A.L., Albert R. Emergence of scaling in random networks // science. – 1999. – Т. 286. – №. 5439. – С. 509-512.
11. Newman M.E.J. Clustering and preferential attachment in growing networks // Physical review E. – 2001. – Т. 64. – № 2. – С. 25-102.
12. Краснов Ф.В., Докука С.В. Моделирование и оценка влияния от применения каркаса Scrum в процессе написания научных статей // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 9, №6 (2017) https://naukovedenie.ru/PDF/17TVN617.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

13. Liben-Nowell D., Kleinberg J. The link-prediction problem for social networks // journal of the Association for Information Science and Technology. – 2007. – Т. 58. – № 7. – С. 1019-1031.
14. Makarov, I., Bulanov, O., Zhukov, L.E. Co-author recommender system // In International Conference on Network Analysis, Springer. – 2016. – С. 251-257.
15. ChamLü L., Zhou T. Link prediction in complex networks: A survey // Physica A: statistical mechanics and its applications. – 2011. – Т. 390. – № 6. – С. 1150-1170.
16. Краснов Ф.В. Модель процесса публикаций научно-практических статей по специальности 25.00 «Науки о Земле» // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 9, №5 (2017) https://naukovedenie.ru/PDF/62TVN517.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.
17. Blei D.M. Probabilistic topic models // Communications of the ACM. – 2012. – Т. 55. – № 4. – С. 77-84.
18. Ketchen Jr.D.J., Shook C.L. The application of cluster analysis in strategic management research: an analysis and critique // Strategic management journal. – 1996. – С. 441-458.
19. Grover, A., Leskovec, J. node2vec: Scalable feature learning for networks // In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. – 2016. – С. 855-864.
20. Makarov I., Bulanov O., Gerasimova O., Meshcheryakova N., Karpov I., Zhukov L. E. Scientific Matchmaker: Collaborator Recommender System // In: van der Aalst W. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2017. Lecture Notes in Computer Science, Springer. – 2018. – Т. 10716, С. 404-410.

Krasnov Fedor Vladimirovich
LLC "Gazpromneft NTC", Saint-Petersburg, Russia
E-mail: Krasnov.FV@Gazprom-Neft.ru

Makarov Ilya Andreevich
National research university «Higher school of economics», Moscow, Russia
E-mail: iamakarov@hse.ru

Predicting co-author relationship for science and technology center of Gazpromneft based on the graph modeling

Abstract. A collective co-authorship of scientific articles has deterministic and random structural components. In addition to the rational aspects of the team assembling of co-authors of the scientific article there are also emotional components. Over time add up and disintegrate the working groups of researchers, updated the organization workforce and the composition of the contractors that take part in joint industry collaborations for conducting research.

Despite the complexity of co-authorship nature, there are several classes of models to simulate the formation of co-authorship. Among them, models based on random graphs and models of the formation of collaborations based on the competencies of the authors. Both the mathematical apparatus developed and used for several decades separately. However, practical applications of the models of collaborations in the corporate practice not so much.

The authors put forward the hypothesis that it is necessary to combine several different types of models to have better understand the nature of scientific collaborations in a separate organization.

The authors of this study set the task to develop a method of constructing a model of co-authorship for scientific-technical center, including various structural components of joint authorship.

As a result, the authors developed a model using machine learning methods, random graphs and models of competencies. Based on the developed model, the forecast for the development of co-authorship in the writing of scientific articles scientific and technical center of GazpromNeft.

The practical value of this study is the following:

1. Quantified the contribution of different structural components in forming collaborations when writing scientific articles.
2. Forecasting the development of co-authorship in the writing of scientific articles allows planning enterprise resources to support the growth of scientific publications.

Understanding the cluster structure of co-authorship makes it possible to align the lines of scientific activity in accordance with the strategic plan for the development of the scientific and technical center.

Keywords: mathematical model of organization process; collaboration graph; graph metrics; node forecasting; machine learning