

# Метрики для раннего прогнозирования успешности R&D

**Ковалев Сергей Петрович**

Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации (Москва)  
Заведующий лабораторией информационных технологий в управлении Института прикладных экономических исследований РАНХиГС  
Заслуженный экономист РФ, кандидат технических наук  
KovalevSePe@yandex.ru

## РЕФЕРАТ

Описан новый математический подход к прогнозированию перспективности исследований и разработок (R&D) в сфере науки и технологий. На выборке текстов, описывающих известные R&D за период с 1955 по 2014 гг., показано, что вероятность успешного прогноза традиционными методами на доинвестиционной фазе жизненного цикла R&D является низкой. Проведена аналогия между подходом к оценке качества текстов компьютерных программ и формализованными научно-техническими текстами, предложены четыре размерно-ориентированные и четыре функционально-ориентированные числовые метрики, основанные на лексическом анализе и эволюционно-онтологическом подходе к выборке текстов R&D. Выведены формулы для расчета размерно-ориентированных метрик, получаемые в результате расчета значения, интерпретированы с точки зрения задачи прогнозирования успешности R&D. Затронуты вопросы построения исходных данных для расчета более сложных функционально-ориентированных метрик с использованием лексико-графовых моделей текстов R&D, решения задач декомпозиции и поиска на графах коллокаций терминов в целях исследования эволюции терминологии, наличия тавтологических определений, оценки качественной структуры текста. Строго математически специфицированы исходные данные, необходимые для расчета предложенных автором восьми формул, используемых для вычисления метрик. Рассчитаны индексы нелинейной парной корреляции метрик и известных результатов R&D на тестовой выборке, представлены результаты исследования вероятности корректного прогноза с помощью предложенных метрик, демонстрирующие, что все они коррелируют с успешностью R&D. Интерпретированы диапазоны получаемых значений для всех описанных метрик, объяснено поведение полученных зависимостей индексов корреляции для метрик и вероятностей успешного прогноза инвестиционной перспективности R&D. Показано, что в сочетании с другими методами применение разработанного математического подхода на основе вычисления восьми числовых метрик может повысить вероятность точного прогноза при оценке целесообразности инвестиций и выборе перспективных направлений исследований и разработок.

## КЛЮЧЕВЫЕ СЛОВА

раннее прогнозирование, перспективные исследования и разработки, R&D, числовые метрики, размерно-ориентированные метрики, функционально-ориентированные метрики, лексический подход, композитные модели оценки R&D

Kovalev S. P.

## The Metrics for Promising R&D Early Forecast

**Kovalev Sergey Petrovich**

The Russian Presidential Academy of National Economy and Public Administration (Moscow, Russian Federation)  
Head of the Laboratory of information techniques of management and control, RANEP Institute of applied economy  
Honoured economist of Russian Federation, PhD in Technical Science  
KovalevSePe@yandex.ru

## ABSTRACT

The paper is concerned to the new methodological approach for early prognosis of promising research and development (R&D) in the fields of science and technology. As it's demonstrated with

the texts' corpus covering the known R&D stories from 1955 to 2014, the correct prognosis probability at the R&D life pre-investment cycle phase based on the traditional econometric methods is not high. The analogy is drawn to computer program text quality metrics based test approach and formally well-structured scientific & engineering texts. Making a start from this analogy, the four size-based quantitative metrics and four functional-based ones basing on lexical approach completed with ontological evolutionary approach to R&D texts' corpus investigations are worked out. The relevant formulas are deduced to calculate the size-based metrics. The resulting values are interpreted from the point of view of promising R&D search and prognosis task. The key questions are described in details for source data formation to calculate more complex functional-based metrics using some lexical-graph R&D text models, to solve decomposition tasks and path search on graphs of terms collocations and co-words with the purpose of terminology evolution investigations, tautological definitions localization, and texts structure quality estimation. The source data necessary for the eight deduced formulas of author's metrics calculation are rigorously specified. The non-linear pair correlation indexes are evaluated for every metric and known R&D historical result on the test text corpus. The probabilities of correct forecast with the eight metrics demonstrate good level of correlation with successful R&D stories. The ranges of resulting values for all the metrics are rigorously described and interpreted, their details of correlation indexes behavior and correct forecast probabilities are explained to support good decision regarding the most promising R&D choice and fulfill a purpose of investment activity at the early phase of R&D life cycle. As it's demonstrated by the author the implementation of described mathematical approach based on the eight metrics results in higher probability of prognosis for better R&D choice and lets an investment manager to achieve the purpose of optimal funding in combination with other known methods.

#### KEYWORDS

early forecast, promising research and development, R&D, quantitative metrics, size-based metrics, functional quality metrics, lexical approach, composed models for promising R&D forecast

«Вы не можете управлять тем, что не можете измерить».

*Tom DeMarco*

## Введение

Несмотря на успехи экспертных, наукометрических и экономико-математических подходов к прогнозированию, предсказание успешности нового продукта по-прежнему остается задачей с высокой степенью неопределенности [7, р. 420–421]. Между тем, процессы разработки стратегии требуют достаточно высокой точности оценки исходного уровня объекта стратегирования. Точность измерений исходного уровня предопределяет вектор и общую кинематику стратегии [9, р. 90]. Традиционные маркетинговые исследования, предшествующие инвестиционным решениям, работают только на поздних стадиях существования научных исследований и разработок (Research and Development, R&D), когда уже существуют образцы продуктов новых технологий, на рынке есть хотя бы приблизительные их аналоги, и можно проводить стоимостный анализ, оценивать соотношение затрат на маркетинг и потенциала рынка. Оценка перспективы R&D, когда они существуют лишь на стадии доказательства жизнеспособности принципов и разработки технологии, представляет проблему, не имеющую строгих решений.

На рис. 1 представлены результаты проведенного в 2016 г. коллективом Лаборатории информационных технологий в управлении РАНХиГС анализа 843 успешных и провальных<sup>1</sup> (с точки зрения доведения до стадии рыночных продуктов) R&D за период с 1955 по 2014 гг. На диаграмме отображены средние значения вероятности рыночной успешности технологии или готовых продуктов на ее основе для трех укрупненных стадий жизненного цикла R&D «идея — технология — продукт».

<sup>1</sup> На основе анализа истории отечественных и зарубежных изобретений и проектов в сфере точных, инженерных и гуманитарных наук и технологий, на базе РГБ и ГПНТБ за период с 1955 по 2014 гг.

Эти вероятности рассчитывались на основе элементарного соотношения:

$$P_i = N_i^+ / N_i^{\text{общ}} \quad (1)$$

где  $N_i$  — число положительных примеров, демонстрирующих уверенное предсказание успеха, отраженное в текстах статей, публикаций, маркетинговых исследований и т. п. за исторический период жизни изобретений и проектов;  $N_i^{\text{общ}}$  — общее число изученных примеров;  $i = \{1, 2, 3\}$  фаза жизненного цикла R&D, для которой исследовались письменные источники.

Как видно из рис. 1, даже для технологической фазы жизненного цикла вероятность точного прогноза успешности составила всего 1/2. Это означает, что с тем же успехом привлеченные эксперты, авторы аналитических и маркетинговых исследований могли бы бросать в воздух монетку. На самой ранней фазе, где основной методикой являются экспертные оценки, вероятность еще ниже, 1/5. То есть, лишь одна из пяти предсказанных историй успеха реализовалась в жизни. Более или менее хороший результат с вероятностью 0,6, т. е. выше случайного броска монетки, наблюдается на стадии вывода на рынок относительно готовых продуктов новых технологий, когда становится возможной финансовая оценка.

Этапы жизненного цикла от постановки научной проблемы до создания технологии и продуктов на ее основе:

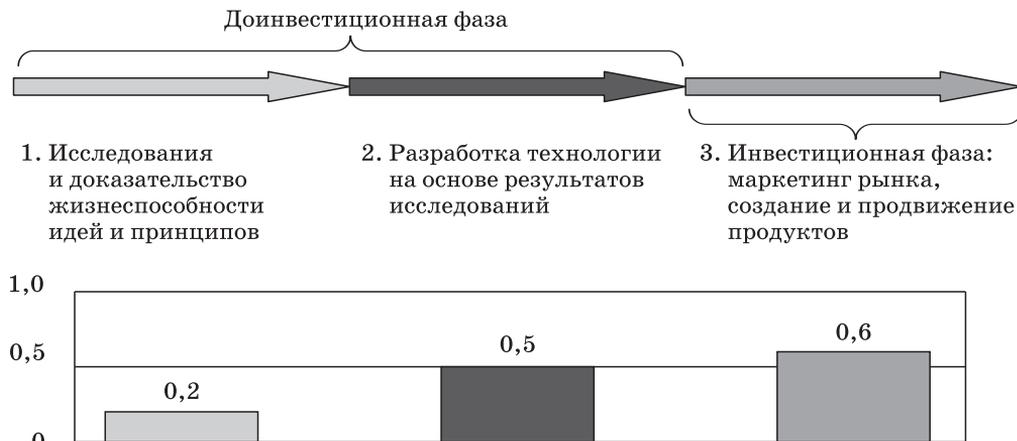


Рис. 1. Вероятность точного предсказания рыночной успешности R&D на трех основных стадиях жизненного цикла на основе традиционных методов оценки

Такой результат нельзя не учитывать в связи с приостановкой работы сайта БД РНТД Минобрнауки России и вводом в эксплуатацию с 1 января 2014 г. Единой государственной информационной системы (ЕГИС) учета научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения, для исполнителей по таким работам<sup>1</sup>.

Появление различных вариантов сетевых экспертно-аналитических систем для прогноза R&D, опирающихся на информационную базу ЕГИС, даже таких авторитетных разработчиков, как ИПУ РАН [3, с. 121–122], также вряд ли радикаль-

<sup>1</sup> Постановление Правительства РФ № 327 от 12.04.2013 «О единой государственной информационной системе учета научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения». С. 2–5.

но улучшит ситуацию, так как фактор субъективизма экспертов (в силу неполноты информации или иных причин) сохраняется независимо от способа, которым получают пул экспертных оценок. Будь то заключения экспертов на бумаге 50-х годов прошлого века, или отправленные ими в сеть сообщения со своего смартфона. Современные информационно-коммуникационные технологии (ИКТ), безусловно, упрощают доступ к информации, обмен экспертными мнениями и статистическую обработку заключений, несколько снижая риск субъективизма любого происхождения [1, с. 34–38], но не меняют саму природу экспертного подхода к прогнозированию. В частности, не снимают проблемы этичности привлечения к экспертизе представителей противоборствующих школ, возможности аффилированности, иной заинтересованности. Альтернативой является применение математических численных методов для оценки объективных характеристик материалов R&D (Mathematical and Quantitative Methods в терминологии JEL Classification System).

### Математический численный метод на основе метрик

Предложим вариант математического численного метода для прогнозирования успешности R&D на доинвестиционных стадиях их жизненного цикла, опираясь на достижения в области разработки программного обеспечения.

В последнее десятилетие в оценке качества и, особенно, в управлении разработкой программного обеспечения произошла настоящая революция, основанная на применении метрик (software metric), т. е. математически вычисляемых мер, позволяющих получить численное значение свойства программного обеспечения или его спецификации.

Один из основоположников этого направления в теории и практике информатики, Том ДеМарко (Tom DeMarco), крылатая фраза которого на конференции 1983 г. [8, р. 753–754] вынесена в эпиграф статьи, предложил рассматривать так называемые размерно-ориентированные и функционально-ориентированные метрики, представляющие собой статистические характеристики текста (кода) программы. Например, количество строк кода, число циклов, количество процедур и т. п. — для первой группы метрик. Алгоритмическую сложность кода, общее количество и иерархию классов, связность<sup>1</sup>, покрытие кода программы тестовыми процедурами, число ошибок и др. — для второй.

Нельзя не отметить, что сам ДеМарко в своих работах не забывает говорить о потенциальной неточности любых мер подобного рода: нет метрик, которые бы одновременно что-то значили и были достаточно точны. В частности, количество строк кода программы — это просто объем текста на некотором языке программирования, и не дает исчерпывающего представления о сложности задачи, решаемой программой. Разумеется, сказанное относится и к предлагаемому далее методу, основанному на прямой аналогии между текстом программы, написанном на современном языке программирования высокого уровня (например, C# или Java) и научно-техническим либо деловым текстом, для которых также свойственны особые структуры документа, обороты и формулы речи.

Отталкиваясь от этой аналогии с метрическим подходом к оценке программного обеспечения, приведем примеры значимых и одновременно простых для вычисления размерно-ориентированных и функционально-ориентированных метрик, подходящих для оценки текстов R&D.

<sup>1</sup> Связность, сцепление (от английского термина cohesion) — мера внутренней взаимосвязи между частями одного модуля сложного программного обеспечения.

**Размерно-ориентированные метрики**

$$\rho = N_f / N_w, \quad (2)$$

метрика «плотности» формул и поясняющих формализованных иллюстраций (диаграмм, схем, графиков) как отношение их числа  $N_f$  к общему объему текстов, выражаемому числом слов во всех представленных для анализа текста. Сама по себе эта метрика не может служить надежным критерием, поскольку свидетельствует лишь об объеме описательной части R&D. Вместе с тем, если документация слишком краткая, это может свидетельствовать о слабой проработке идеи исследования. И наоборот, излишние словесные описания при малом числе графических иллюстраций и математических формулировок могут (но не обязаны) говорить о вненаучности тематики R&D.

$$\mu = \begin{cases} 1,0, & \max(L_p) < 800 \\ 0,5, & \max(L_p) \in [800, 1024] \\ 0,0, & \max(L_p) > 1024, \end{cases} \quad (3)$$

нечеткая (fuzzy) метрика максимальной длины абзаца  $L_p$ , косвенно характеризующая структурированность текста. Если в программировании сегодня считается хорошим стилем, когда код отдельной процедуры или класса не превышает 20–30 строк, то для текстов R&D такой четкой грани нет. В рассматриваемом случае, учитывая ГОСТ для документации, общие правила научного стиля изложения, можно принять, что каждый отдельный абзац, выражающий определенную мысль или несколько последовательных заключений, не должен превышать половину страницы А4, типичный размер которой для шрифта 14 через полтора интервала составляет 1400–1600 знаков. Максимальное значение метрики 1.0 соответствует условно хорошему стилю изложения.

В качестве еще одной метрики структурированности текста предложим среднее квадратичное отклонение длины абзаца, которая может косвенно свидетельствовать о равномерности и непрерывности изложения материала

$$\sigma = \sqrt{\frac{1}{n} \sum_{p=1}^n (L_p - \bar{L}_p)^2}, \quad (4)$$

где  $n$  — общее число абзацев в тексте;  $L_p$  — длина абзаца;  $\bar{L}_p$  — средняя длина абзаца  $p$ . Чем больше отклонение, тем больше неравномерность в изложении материала.

С двумя последними метриками тесно связана еще одна предлагаемая нами величина:

$$\theta = N_h / N_p, \quad (5)$$

отношение числа заголовков в тексте к общему числу абзацев, косвенно показывающее, насколько системно построено изложение материала.

При простоте соотношений (2)–(5), как показал результат нашего исследования вероятности корректного прогноза (рис. 1), все четыре предложенные метрики в разной степени коррелируют с успешностью R&D.

На рис. 2 представлены индексы нелинейной парной корреляции  $R_{xy}$ , вычисленные методом регрессионного анализа с помощью статистического пакета SPSS™ на упомянутой во втором абзаце статьи выборке из 843 описаний R&D.

Как видно из диаграммы (рис. 2), прослеживается устойчивая, хотя и не слишком сильная, связь между метриками и успешностью R&D, анализируемой на основе своих текстовых описаний.

Нельзя не отметить некоторую парадоксальность поведения метрик в зависимости от фазы жизненного цикла R&D, прослеживаемую на рис. 2 с помощью

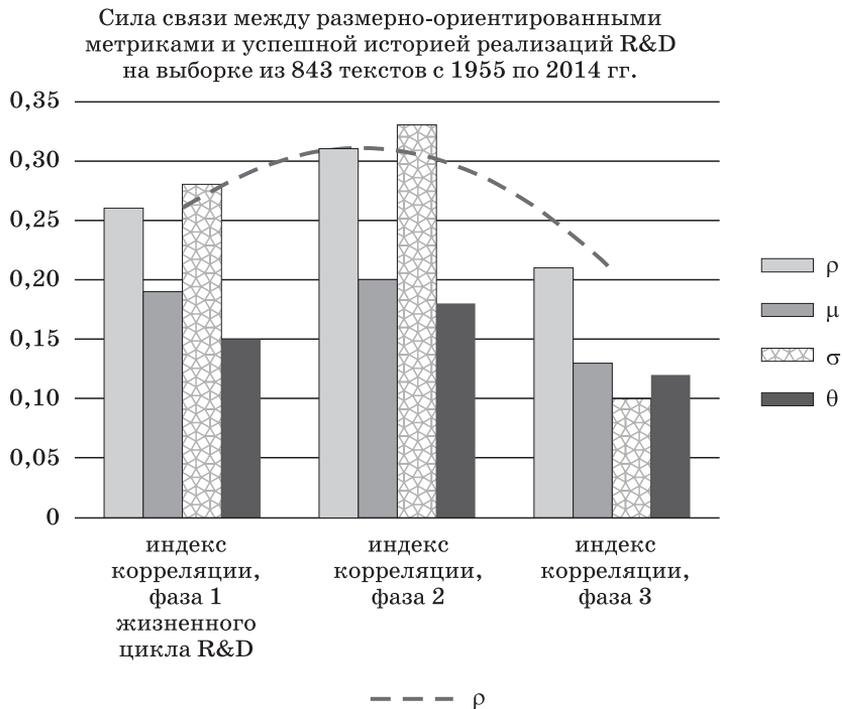


Рис. 2. Индексы нелинейной корреляции, иллюстрирующие силу связи между предложенными метриками и успешностью прогноза для трех фаз жизненного цикла R&D

линии полиномиальной аппроксимации тренда. Максимальная сила связи для прогноза на фазе разработки технологии снижается и становится минимальной к инвестиционной фазе жизненного цикла R&D. Это поведение мы можем объяснить появлением на последней фазе большого числа коммерческих и рекламных текстов, не стесненных, естественно, жесткими рамками стилистики текстов научных описаний и технологической документации. Вместе с тем, на доинвестиционных фазах связь достигает наибольших значений, что позволяет говорить о применимости четырех предложенных размерно-ориентированных метрик именно для раннего прогнозирования успешности R&D.

Для построения более надежного прогноза необходимо дополнять вычисление размерно-ориентированных метрик анализом функциональных характеристик R&D.

### Функционально-ориентированные метрики

Набор значимых для целей прогнозирования успешности R&D функционально-ориентированных метрик значительно сложнее. Будем строить их на базе онтологического подхода к исследованию эволюции предметной области [4, с. 378–379], заключающегося в следующем.

Каждой тематике R&D сопутствуют свои термины и их сочетания (коллокации, co-words, collocations). «Рождение» любого научно-технологического направления фиксируется в момент появления принципиально нового понятия, которое можно назвать корневым. Развитие и углубление исследований отражается в языке текстов

в виде своего рода эволюции производных терминов и их устойчивых словосочетаний — коллокаций. Формализованные тексты также ссылаются друг на друга, на публикации предшественников и основоположников направления, фундаментальные труды по дисциплинам, положенным в основу R&D и т. д.

С помощью современных web-поисковых технологий можно находить коллокации терминов и связанные между собой ссылками тексты по одной тематике или смежным с ней. Будем рассматривать каждый такой термин или коллокацию как вершину математического графа, а текст, в котором они появляются, как вершину гиперграфа.

Отметим особо, что при анализе текстов не делается попыток решения лингвистических задач определения словообразования, синтаксиса, морфологии слов и т. п., требующих сложнейших алгоритмов и до конца еще никем не решенных. Все слова и словосочетания берутся в той форме, в какой они появляются в тексте, как завершённые «лексические единицы» речи (lexical chunks) в рамках подхода, предложенного в 1993 г. Майклом Льюисом (Michael Lewis) для быстрого обучения английскому языку и развитого впоследствии в стройную систему [6, р. 128–130]. Фактически этот же подход сегодня используют WEB-поисковые системы (Google, Bing, Yandex и пр.), накопившие огромные устойчивые словари словосочетаний, встречающихся в поисковой строке.

Построив граф терминов и связей между ними, можно наглядно представить картину эволюции текстов, посвященных анализируемой R&D.

Проводя известными способами разбиение данного графа на связанные области [2, с. 7], вычисляя степень связанности текстов между собой, находя циклические ссылки между терминами, строя кратчайшие пути, выполняя другие подобные хорошо изученные операции на графах — можно решать интеллектуальные задачи оценки качества и перспективности R&D, не прибегая к интеллектуальному потенциалу экспертов.

В табл. 1 приведены четыре типовые задачи на графах, решение которых дает характеристики текстов R&D, пригодные для последующего построения четырех функционально-ориентированных метрик. Самая первая задача, которую решает любой эксперт, изучающий предъявленный ему текст, — это определение (локализация) предметной области исследования в соответствии с одной из принятых классификаций (УДК, JEL и т. п.). Построение покрывающего дерева на графе терминов — это нахождение последовательности всех вершин графа, не содержащей циклов, начиная от выбранной вершины, называемой корнем дерева. В общем случае можно построить несколько вариантов такого дерева. В рассматриваемом случае последовательности вершин-терминов образуют терминологический словарь, который можно затем сравнивать, например, с профессиональными online-словарями (медицинскими словарями, экономическими, инженерными и т. д.), определяя принадлежность R&D к уже известной предметной области науки и технологии. Или наоборот, выявляя полную новизну использованной авторами R&D терминологии.

Вторая задача, выявление признанного научно-методологического фундамента R&D (или его отсутствие), решается в ходе нахождения путей на графе связей от вершины — ключевого термина R&D к основополагающим терминам, вплоть до корневого понятия, с которого началось развитие нового направления в науке и технологии. Например, в случае с использованным нами «лексическим подходом» цепочка ссылок в найденных Google текстах через несколько шагов приводит к корневому термину «lexical chunk» в вышеупомянутой основополагающей работе М. Льюиса.

Решение задачи нахождения циклов на графе связи терминов и опубликованных текстов позволяет выявить тавтологии, т. е. ссылки определений самих на себя

**Сопоставление традиционных задач на графе происхождения и связи терминов с интеллектуальными задачами анализа R&D**

Типовая задача на графах, для которой есть вычислительные алгоритмы или эвристики	Интеллектуальная задача функционального анализа R&D
Построение покрывающих деревьев	Локализация предметной области исследования и создание терминологического словаря
Нахождение путей от целевой вершины — термина к корневому (начальному) термину	Доказательство научно-методологической обоснованности
Нахождение циклов, проходящих через определенную вершину и корневой термин	Обнаружение тавтологических определений и циклических ссылок
Кластеризация, раскраска вершин	Выявление смежных по тематике направлений R&D, не связанных между собой

через несколько шагов. Чаще всего тавтологические определения свидетельствуют о вненаучности подхода, использованного авторами R&D, возможно даже псевдонаучности заявленной темы.

Наконец последняя задача, кластеризация вершин гиперграфа текстов, поставляет данные для нахождения похожих между собой, но связанных с различными авторскими коллективами тематик R&D. Эти данные можно использовать при построении функционально-ориентированной метрики для численной оценки степени конкуренции, существующей в анализируемом направлении R&D.

Прежде чем перейти к описанию функционально-ориентированных метрик, вычисляемых на основе результатов решения перечисленных задач на графе связей терминов и текстов R&D, специфицируем получаемые данные в форме, пригодной для дальнейших математических расчетов.

(1) Локализованный терминологический словарь предметной области R&D представляет собой частично-упорядоченное множество  $U_v$  лексических единиц в смысле Льюиса. Его пересечение с одним из профессиональных терминологических словарей также представляет собой множество  $U_c^i = U_v \cap U_s^i$ . Если пересечение пусто,  $U_c^i = \emptyset$ , то R&D не принадлежит предметной области  $i$ . Если пересечение тождественно равно  $U_v$ ,  $U_c^i \equiv U_v$ , это означает полное соответствие тематики R&D известной научной дисциплине. Отношение мощности пересечения к мощ-

ности исходного множества,  $\frac{\|U_c^i\|}{\|U_v\|} \in [0, 1]$ , показывает степень связи предметной

области анализируемой R&D и одной из известных науке дисциплин.

(2) Путь  $\{T^m, T^{m-1}, \dots, T^0\}$  от одного из ключевых терминов, используемых авторами для описания идей и планируемых результатов своих R&D, к строго определенному научному термину, содержащий  $m - 1$  промежуточных вершин-терминов, в случае своего существования, показывает опосредованную связь с устоявшейся научной методологией исследования. Длина пути  $m$  может свидетельствовать о степени такой связи, а отношение числа путей  $N_p$  и терминов  $N_T$  с учетом длины путей — об опоре исследования на доказанный научно-методологический базис.

- (3) Цикл  $C_j = \{T_j^0, T_j^m, T_j^{m-1}, \dots, T_j^1, T_j^0\}$ , проходящий через вершину графа, соответствующую ключевому термину  $T_j^0$ , встречающемуся в описаниях R&D, означает тавтологическое определение. Соотношение числа  $N_c$  таких циклов к числу ключевых терминов  $N_T$ , в которых авторы описывают свои исследования и разработки, может говорить о степени научности (вплоть до псевдонаучности, если тавтологическими являются все термины и определения, не связанные с научными дисциплинами).
- (4) Число  $N_K$  кластеров, характеризующих сходные между собой тематические исследования различных авторских коллективов в сочетании с мощностью конкретного кластера текстов  $\|K_i\|$ , выявленного для анализируемого описания R&D, могут указывать на степень оригинальности проводимых исследований и разработок, уровень конкуренции, сложившейся в данном направлении.

Отметим, что получение перечисленных данных требует решения ряда дополнительных web-поисковых, статистических и аналитических задач, что может достигаться использованием расширенных многоагентных моделей для построения релевантной выборки и дальнейшего исследования корпуса текстов R&D лексическими методами [5, с. 114–115].

Основываясь на четырех видах специфицированных данных, перейдем к описанию собственно функционально-ориентированных метрик.

$$\varepsilon = \frac{1}{n} \sum_{i=1}^n \frac{\|U_C^i\|}{\|U_V\|}, \quad (6)$$

метрика общей связи с признанными научными дисциплинами и методологиями, вычисление которой дает дробное число от 0, что соответствует венаучности текстов R&D, до 1 — абсолютная опора на предшествующие результаты апробированных научных исследований. Заметим, последнее отнюдь не всегда хорошо, поскольку может свидетельствовать о вторичности, отсутствии оригинальности темы R&D. Каково оптимальное значение — вопрос дискуссионный, требующий дополнительных статистических исследований; к тому же с очевидностью зависящий от предметной области R&D. В своей работе мы полагаем его приемлемым в диапазоне  $\varepsilon \in [0,5, 0,7]$ .

Эта метрика дополняется метрикой терминологической связности:

$$\tau = \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{m^i}{m_{\max}}, \quad m_{\max} > 0, \quad (7)$$

$$0, \quad m_{\max} = 0$$

где  $m^i$  — длина пути от ключевого термина R&D к корневому научному термину, до которого прослеживается его родословная на терминологическом графе. Если для какого-либо термина такого пути построить не удалось, то значение принимается 0. Значение длины максимального пути  $m_{\max}$  равно 0, если нет ни одного пути, хотя от одного ключевого термина, хотя к одному корневому.

Как и для предыдущей метрики, значение 1 не всегда соответствует максимальной вероятности успешного прогноза для анализируемых R&D, так как может (но не обязано) попросту означать отсутствие собственных оригинальных идей у авторов R&D. Аналогично предыдущему, полагаем его значение приемлемым в диапазоне  $\tau \in [0,5, 0,7]$ , оставляя вопрос открытым для уточнения в ходе дальнейших исследований.

$$\varphi = 1 - \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{N_c^i}{N_c}, \quad N_c > 0; \quad (8)$$

$$1, \quad N_c = 0$$

где  $\sum_{i=1}^{N_T} N_C^i$  — общее число найденных циклов, проходящих через все вершины-термины. Если для какого-либо термина цикла не найдено, то значение  $N_C^i$  принимается равным 0. Если  $N_C = 0$ , т. е. не найдено ни одного цикла (тавтологического определения), то метрика принимает свое максимальное значение 1. Любой найденный цикл уменьшает это значение, вплоть до 0, что может свидетельствовать об отсутствии корректно выстроенной аксиоматики для теорий, на которые авторы ссылаются в текстах R&D. Эту метрику можно назвать метрикой аксиоматичности. Метрика степени оригинальности и конкурентности R&D

$$\omega = 1 - \frac{1}{N_K} \sum_{i=1}^{N_K} \frac{\|K_i\|}{\|K\|}, \tag{9}$$

где  $\|K\| = \max(\|K_i\|)$  максимальная мощность среди всех  $N_K$  найденных кластеров мощности  $\|K_i\|$  каждый, принимает значения в диапазоне от 0 до 1. Чем больше число кластеров — конкурентов анализируемой R&D, и чем они мощнее, тем ближе значение метрики к 0, показывая, что исследование не оригинально, и у него много конкурентов. Если R&D оригинальна и значение метрики 0, то следует более тщательно вычислять и анализировать предыдущие три метрики, характеризующие научность проводимых исследований и разработок.

Приведем диаграмму степени связи между четырьмя предложенными функционально-ориентированными метриками и вероятностью успешного прогнозирования (рис. 3), построенную на нашей выборке (см. также рис. 1 и 2).

Сила связи между функционально-ориентированными метриками и успешной историей реализаций R&D на выборке из 843 текстов с 1955 по 2014 гг.

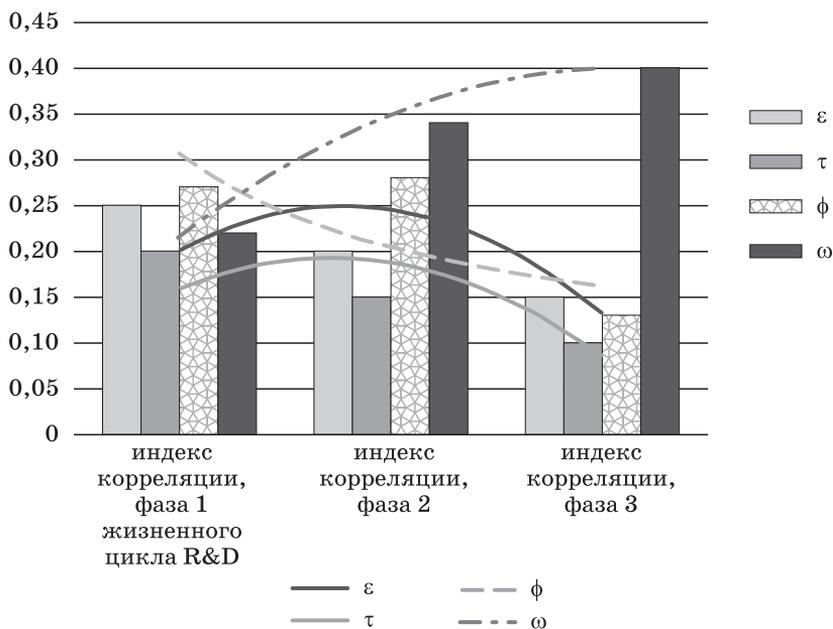


Рис. 3. Индексы нелинейной корреляции, иллюстрирующие силу связи между функционально-ориентированными метриками и успешностью прогноза для трех фаз жизненного цикла R&D

Нанесенные на диаграмме линии полиномиальной аппроксимации трендов для каждой из исследуемых метрик показывают, что поведение функционально-ориентированных метрик значительно сложнее, чем описанных ранее размерно-ориентированных.

Наибольшая корреляция для ранних фаз R&D наблюдается у метрик  $\varepsilon$  общей связи с основополагающими научными дисциплинами и терминологической связности  $\tau$ .

Метрика аксиоматичности  $\varphi$ , оценивающая качество определений, даваемых авторами ключевым терминам своих исследований и разработок, демонстрирует удовлетворительные значения индекса корреляции для исследовательской и технологической фазы жизненного цикла R&D. Она снижается к фазе создания и продвижения коммерческого продукта. Как и в случае с размерно-ориентированными метриками, мы можем объяснить этот эффект преваляцией рекламных текстов над документацией.

В этом смысле логично выглядит также рост индекса корреляции между метрикой оригинальности и конкурентности  $\omega$  и вероятностью успеха R&D: чем более оригинальной является проделанная работа, доведенная до стадии рыночного продукта, тем выше вероятность успеха и дальнейшей жизни практических результатов R&D.

## Заключение

Опираясь на аналогию между современным подходом к оценке качества программного обеспечения, предложены восемь математически строго вычисляемых характеристик — метрик для текстов R&D, позволяющие решать задачу прогнозирования успешности R&D на ранних фазах их жизненного цикла «идея—технология—рыночный продукт».

На значимой выборке из 843 историй успеха R&D, изученных на основании имеющихся в библиотечных фондах описаний за период с 1955 по 2014 гг., подтверждена устойчивая связь между вероятностью прогноза и размерно-ориентированными метриками, характеризующаяся индексом нелинейной корреляции не ниже 0,15 для ранних фаз и 0,10 для рыночной фазы R&D. На основе лексического подхода к анализу текстов R&D, находимых в электронном виде web-поисковыми системами, с применением аппарата математической теории графов, предложены четыре функционально-ориентированные метрики с индексами корреляции не ниже 0,20 для начальных фаз жизненного цикла успешных R&D.

Показано, что демонстрируемое всеми метриками поведение логически объяснимо характером изменения корпуса анализируемых текстов от фазы к фазе.

В заключение изложения еще раз отметим, что любая по отдельности взятая метрика не может служить надежным индикатором для прогнозирования успешности R&D. Только совместное использование всех восьми предложенных взаимодополняющих размерно-ориентированных и функционально-ориентированных метрик дает возможность значимого прогноза, на который может опираться инвестиционный менеджер в процессе принятия решений о направлениях и объемах финансирования перспективных исследований и разработок. При этом мы никоим образом не считаем приведенные результаты нашей работы исчерпывающими, и полагаем, что описанная методология способна дать хороший практический результат в первую очередь при ее разумном сочетании с традиционными экономико-математическими моделями и методами прогнозирования.

## Литература

1. Губанов Д. А., Коргин Н. А., Новиков Д. А., Райков А. Н. Сетевая экспертиза. 2-е изд. / под ред. чл.-к. РАН Д. А. Новикова, проф. А. Н. Райкова. М.: Эгвес, 2011. 166 с.

2. Курейчик В. В., Сороколетов П. В. Композитные методы разбиения графовых моделей. Таганрог : изд-во ТРТУ, 2006. 140 с.
3. Райков А. Н. Инновационное саморазвитие сетевой экспертной среды // Организация саморазвивающихся инновационных сред : Сб. статей / под ред. В. Е. Лепского. М. : «Когито-Центр», 2012. С. 120–139.
4. Сороколетов П. В. Композитные модели поиска перспективных R&D на основе мульти-агентных систем // Материалы 9-й конференции «Информационные технологии в управлении» (ИТУ-2016). СПб. : АО «Концерн «ЦНИИ «Электроприбор», 2016. 896 с.
5. Сороколетов П. В. Расширенная многоагентная модель прогнозирования перспективности R&D // Труды конгресса по интеллектуальным системам и информационным технологиям IS&IT'16, 2–9 сентября 2016, Дивноморское, Россия, т. 2, с. 112–117.
6. Lewis M. Language in the lexical approach. In *Teaching Collocation: Further Developments In The Lexical Approach* // Language Teaching Publications, 2000. P. 126–154.
7. Matzler K., Grabher C., Huber J., Füller J. Predicting new product success with prediction markets in online communities // *R&D Management*, Volume 43, Issue 5, November 2013. P. 420–432.
8. DeMarco T. A metric of estimation quality // *AFIPS National Computer Conference*, 1983. P. 753–756.
9. Kvint V. *Strategy for the Global Market*. New York : Routledge, 2016. 518 p.

## References

1. Gubanov D. A., Korgin N. A., Novikov D. A., Raykov A. N. *Network expertize* [Setevaya ekspertiza]. Second edition, ed. by RAS correspondent-member D. A. Novikov, prof. A. N. Raykov. M. : EGVES, 2011. 166 p. (rus)
2. Kureichik V. V., Sorokoletov P. V. *Composed methods for graph models parting* [Kompozitnye metody razbieniya grafovykh modelei]. Taganrog, TRTU, 2006. 140 p. (rus)
3. Raykov A. N. *Innovation self-development of experts' network environment* [Innovatsionnoe samorazvitie setevoi ekspertnoi sredy] // Self-developing innovation environment organization [Organizatsiya samorazvivayushchikhsya innovatsionnykh sred]. The articles collection edited by B. E. Lepsky. M. : «Cognito-Center », 2012. P. 120–139. (rus)
4. Sorokoletov P. V. *Composed models for perspective R&D search based on multiagent systems* [Kompozitnye modeli poiska perspektivnykh R&D na osnove mul'tiagentnykh system] // The 9-th Conference «Information technologies in control» (ITC-2016). SPb. : PLC «Concern «CNII «Electropribor», 2016. 896 p. P. 377–384. (rus)
5. Sorokoletov P. V. *Advanced multiagent model for R&D forecast and estimation* [Rasshirennaya mnogoagentnaya model' prognozirovaniya perspektivnosti R&D] // International Congress on intelligent systems and information techniques IS&IT'16, 2–9 September 2016, Divnomorskoe, Russian Federation, v. 2. P. 112–117. (rus)
6. Lewis M. *Language in the lexical approach*. In *Teaching Collocation: Further Developments In The Lexical Approach* // Language Teaching Publications, 2000. P. 126–154.
7. Matzler K., Grabher C., Huber J., Füller J. *Predicting new product success with prediction markets in online communities* // *R&D Management*, Volume 43, Issue 5, November 2013. P. 420–432.
8. DeMarco T. *A metric of estimation quality* // *AFIPS National Computer Conference*, 1983. P. 753–756.
9. Kvint V. *Strategy for the Global Market*. New York : Routledge, 2016. 518 p.