
**ЭЛЕКТРОННАЯ КОМПОНЕНТНАЯ БАЗА
И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ УПРАВЛЕНИЯ**

УДК 004.048:519.765

**ВЫЯВЛЕНИЕ ПРИЗНАКОВ ИНЖЕНЕРНО-ТЕХНОЛОГИЧЕСКИХ
КОМПЕТЕНЦИЙ ОРГАНИЗАЦИЙ МЕТОДОМ ТЕМАТИЧЕСКОГО
АНАЛИЗА МАССИВА ТЕКСТОВОЙ ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ**

© 2023 г. Е. В. Яковлев^{1,*}, Ю. С. Яковлева²

¹АО «ЦНИИмаш», Королев, Россия

²Военный инновационный технополис «ЭРА», Анапа, Россия

*E-mail: YakovlevEV@tsniimash.ru

Поступила в редакцию 20.11.2023 г.

После доработки 20.11.2023 г.

Принята к публикации 26.11.2023 г.

Рассмотрены аспекты машинной обработки результатов опроса 27 отечественных организаций, проведенного в 2021–2022 гг., применительно к массиву технических документов с описанием инновационных разработок и проектов организаций в текстовом формате. По сравнению с предшественниками впервые рассмотрена проблематика применения метода машинного тематического анализа для обработки текстов с целью определения уникального набора инженерно-технологических компетенций организаций. С помощью принципа распределения Дирихле решалась задача поиска в массиве текстов ключевых слов и словосочетаний, содержание и частотность появления которых позволили распределить их по тематикам. А тематики интерпретировались в качестве признаков наличия у организаций уникальных компетенций. Полученные результаты исследования могут быть использованы специалистами при анализе и мониторинге отраслевых научно-технических направлений, при проведении экспресс-оценки актуальности и новизны инновационных разработок.

DOI: 10.56304/S2782375X23040162

ВВЕДЕНИЕ

В настоящее время в связи с постоянным увеличением объема производимой и хранимой оцифрованной информации, расширением потоков и источников поступления информации, массовости применения приемо-передающих устройств и ростом их функциональности наблюдается повышенный интерес к направлениям автоматизации обработки и структурирования больших массивов информации.

Приоритетной формой поиска технологий и взаимодействия с коммерческими компаниями является непрерывный мониторинг технологий, основной целью которого является настройка аналитической системы циклического выявления сигналов технологической активности отрасли по приоритетным научно-техническим направлениям деятельности организации для сбора и обработки массивов потоковой информации. Осуществление технологического мониторинга направлено на поиск технологий лидеров отрасли или на выявление неявных сигналов и признаков появления новых прорывных идей, уникальных компетенций. Анализ полученной информации позволяет актуализировать корпоративную базу

пилотных проектов, своевременно определять тенденции развития отрасли, проводить технологическую и конкурентную разведку, определять лидеров технологий, прогнозировать новые производственные мощности, инвестировать в технологии опережающего развития.

Таким образом, система знаний о методах сбора, идентификации и анализа информации, основанная на технологических сигналах или признаках разной интенсивности, предоставляет дополнительные возможности и преимущества при исследовании новых технологических аспектов в отрасли.

Актуальность настоящего поискового исследования основана на перспективности машинной обработки и тематического анализа больших массивов текстовой информации по сравнению с работой в ручном режиме на примере анализа текстовых описаний инновационных разработок.

Целью работы является выявление ключевых и второстепенных признаков инженерно-технологических компетенций организаций по направлению “Малые космические аппараты и сопутствующие технологии”.

Следует отметить, что в рамках проведенных обзорных исследований существующих публикаций о применимости методов тематического анализа информации выявлены следующие основные направления применимости данных методов:

- навигация по базам научной и новостной информации [1, 2];
- классификация и категоризация документов [3];
- информационный поиск [4], в том числе многоязычный поиск [5];
- ранжирование информации в рекомендательных системах [6].

Основной сферой применения методов тематического моделирования является социальная сфера: социальные сети, модерирование информационных блогов, оценка эмоционального настроения пользователей интернет ресурсов. При этом применение тематических методов в сфере обработки научной информации развито незначительно по сравнению с социальной сферой [7].

На государственном уровне также отмечена перспективность обработки больших данных [8], которая нацелена на повышение эффективности процессов планирования, прогнозирования и принятия управленческих решений, в том числе за счет увеличения объема доступных данных и развития информационно-коммуникационной инфраструктуры для обеспечения доступа к наборам таких данных.

Впервые рассмотрен тематический анализ текстовых массивов технической информации применительно к выявлению признаков инженерно-технологических компетенций организаций, ранжированию признаков компетенций на ключевые и второстепенные, построению уникальных профилей компетенций организаций.

Основываясь на сказанном выше, необходимо отметить, что представленная работа отражает результаты поисковых исследований по проблематике машинной обработки информации, выявления признаков инженерно-технологических компетенций организаций с использованием метода тематического анализа больших массивов текстовых технических данных.

ОСНОВНАЯ ЧАСТЬ

В рамках данной работы проводили проверку гипотезы о возможности анализа массива (поток) технической текстовой информации на предмет скрытых тематических сигналов или информационных признаков имеющихся в организации инженерно-технологических компетенций по направлению “Малые космические аппараты и сопутствующие технологии”.

В основе гипотезы рассмотрена идея о том, что ключевые слова, технические термины, их синтаксические конструкции и семантическая значимость применимы для выявления скрытых сигналов и принадлежности их к определенной организации. Контекстная логика технических описаний разработок в данном случае использована в качестве основы для выявления спектра признаков инженерно-технологических компетенций с помощью методов машинной обработки корпуса технических документов.

Примененный метод тематического моделирования относится к методам вероятностного статистического анализа больших текстовых данных и предназначен для выявления скрытых тематик в корпусе документов [9, 10].

Большинство современных тематических моделей строится на основе принципа распределения Дирихле [11, 12]. Основным механизмом работы принципа является поиск схожих тематик по целому документу или фрагменту текста. При этом тематическими векторами частот представляются документы, объекты, упоминаемые в документах или связанные с документами термины, названия организаций, годы активности, названия конференций, журналов, статей и другие материалы, что повышает эффективность и гибкость настройки системы обработки информации.

Результаты обработки данных представлены на рис. 1. Половина всех участников опроса относится к малому бизнесу и территориально располагается в Центральном федеральном округе РФ. Представленные разработки относятся к коммерческим проектам (70%). Уровень технологической готовности разработок заявлен как средний (39%) и высокий (32%).

Всего представлена 71 разработка с техническим описанием в текстовом формате на русском языке. Документы в электронном виде структурированы по информационным блокам:

- название проекта;
- контактная информация;
- категория проекта;
- объекты интеллектуальной собственности;
- актуальность разработки;
- описание;
- конкурентные преимущества;
- уровень технологической готовности;
- реализация.

Для формирования исходного датасета применяли алгоритм потокового извлечения текстовых данных на основе объектно-ориентированного языка программирования Python с открытым исходным кодом. Алгоритм позволил значительно ускорить процесс формирования датасета для



Рис. 1. Общие сведения об организациях и разработках: а — привязка организаций к федеральным округам РФ, б — формы собственности организаций, в — разработки коммерческих и некоммерческих организаций, г — уровень технологической готовности разработок [13].

проведения дальнейших исследований и визуализации результатов.

В целях снижения нерелевантных сигналов и информационного шума из датасета исключены неинформативные слова, слова с низким уровнем значимости с частотностью вхождений в описание разработок менее 3. Для повышения уникальности тематик удалены частые слова, которые встречаются в более чем половине описаний всех разработок. К оставшимся словам применен процесс лемматизации с преобразованием слов в словарные формы.

Предварительно обработанные тексты представляют собой списки уникальных слов (токенов), выделенных из строк текста естественного языка. На рис. 2 представлены группировки токенов. Горизонтальная ось — численность группировок, вертикальная ось — количество таких группировок в датасете. Основная часть группировок содержит от 300 до 500 токенов.

Для определения уникальных списков ключевых токенов — тематических ядер — применяли

принцип распределения Дирихле. Полученные таким образом тематические ядра интерпретировали в качестве инженерно-технологических признаков компетенций, которыми потенциально могут обладать организации-участники опроса.

Оптимальное число инженерно-технологических признаков, выделенных из общего контекста технических описаний разработок, определяли по максимальному значению когерентности. Чем выше величина когерентности, тем лучше качественные характеристики текста: семантическая значимость, смысловая логичность, непрерывность. Выявлено, что максимальное значение когерентности 0.43 указывает на возможность структурирования всего массива текстов с выделением 30 уникальных тематических ядер для дальнейшей визуализации и интерпретации признаков инженерно-технологических компетенций.

Визуализацию и интерпретацию осуществляли на основе набора 10–15 ключевых самых частых токенов из общего списка, содержащего 100 токенов (рис. 3) в каждом ядре.

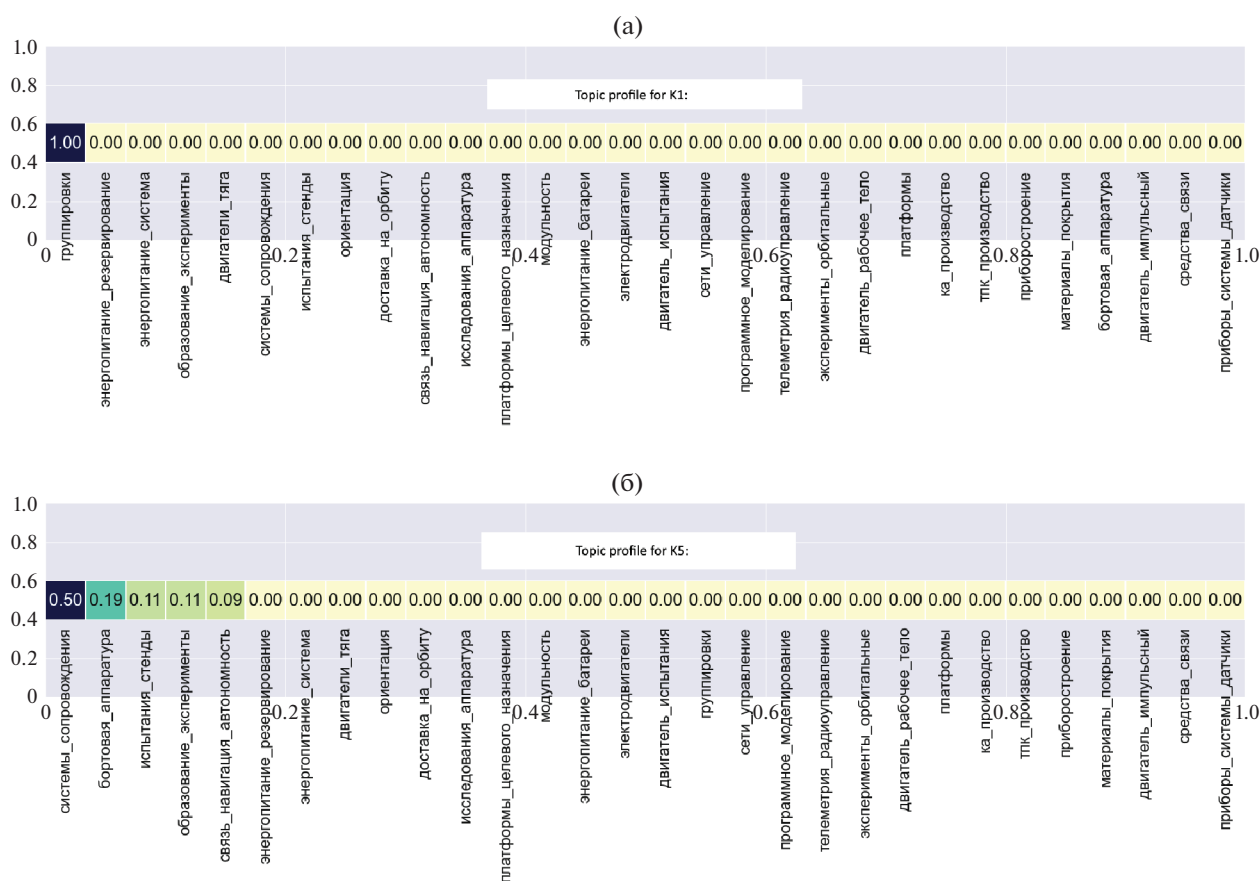


Рис. 4. Примеры индивидуальных профилей признаков инженерно-технологических компетенций организаций и их вероятностная значимость в деятельности компании от 0 до 1: а – организация 1: “Группировки” (1.0), б – организация 5: “Системы сопровождения” (0.5), “Бортовая аппаратура” (0.19), “Стенды и испытания” (0.11), “Образование и эксперименты” (0.11), “Автономные системы связи и навигации” (0.09).

повысить уникальность и интерпретируемость результатов моделирования.

ЗАКЛЮЧЕНИЕ

Впервые рассмотрены вопросы применимости методов тематического анализа для выявления признаков наличия инженерно-технологических компетенций организаций на основе анализа массива текстовых описаний инновационных разработок. Исследование проведено с использованием результатов мониторингового опроса 27 отечественных организаций в 2021–2022 гг. в рамках отраслевого технологического направления “Малые космические аппараты и сопутствующие технологии”.

На основе полученных результатов можно сделать заключение о целесообразности применения методов тематического анализа в отношении текстовых технических описаний инновационных разработок для получения первичной осведомленности о наличии ключевых и второстепенных

признаков инженерно-технологических компетенций.

Полученные результаты исследования могут быть предложены в качестве современного инструмента экспертам, аналитикам, специалистам при автоматизации мониторинга научно-технических направлений, прогнозирования появления новых технологических направлений, оперативного выявления уникальных разработок потенциальных лидеров отрасли, оценки актуальности существующих разработок, оценки перспективности внедрения инновационных разработок в массовое производство.

Проведенная работа носит характер поискового исследования, соответственно, может рассматриваться в качестве основы для дальнейшей проработки направления построения экспертных и рекомендательных систем, в том числе с возможным расширением по следующим направлениям исследований:

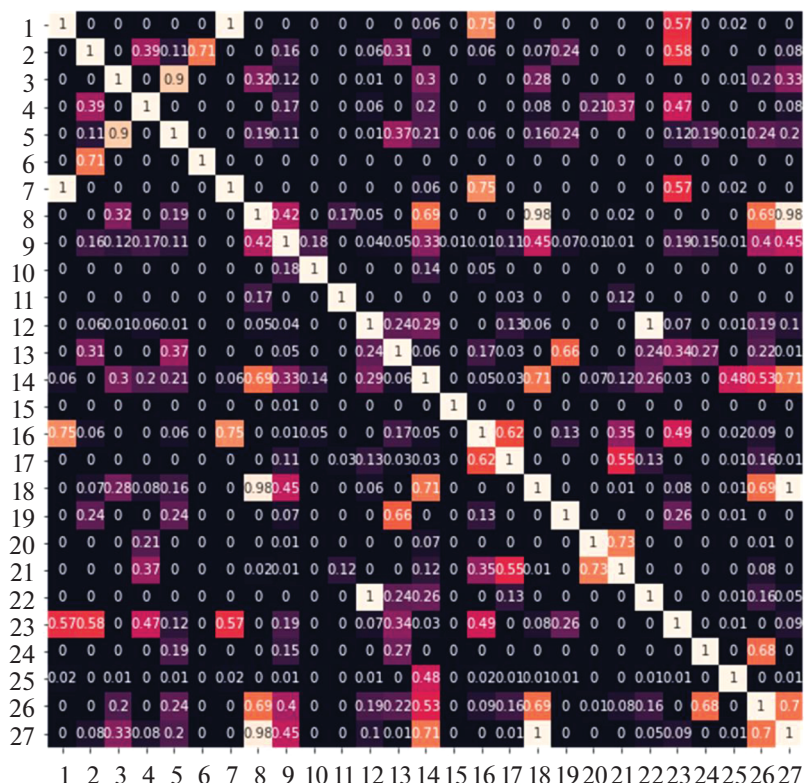


Рис. 5. Тепловая карта соответствия (“похожести”) организаций по признакам инженерно-технологических компетенций: 1 — организации имеют схожие индивидуальные профили, 0 — организации не имеют одинаковых компетенций.

— разработка экспресс-способов поиска, анализа, оценки отраслевых организаций с заданными уникальными компетенциями;

— прогнозирование уровня технологической конкуренции на основе оценки компетенций отраслевых лидеров;

— оценка перспективности индустриальных кооперационных связей, партнерств и консорциумов;

— повышение эффективности и гибкости настройки системы обработки информации за счет использования нескольких источников информации, таких как научные публикации, патентная информация, материалы выставок, конференций, семинаров и др.

СПИСОК ЛИТЕРАТУРЫ

1. Zhang J., Song Y., Zhang C., Liu S. // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.
2. Cui W., Liu S., Tan L. et al. // IEEE Transactions on Visualization and Computer Graphics. 2011. V. 17. № 12. P. 2412.
3. Rubin T.N., Chambers A., Smyth P., Steyvers M. // Mach. Learn. 2012. V. 88. № 1. P. 157.
4. Yi X., Allan J. // Advances in Information Retrieval. Berlin; Heidelberg: Springer, 2009. V. 5478 of Lecture Notes in Computer Science. P. 29.
5. Vulić I., Smet W., Moens M.-F. // Inf. Retr. J. 2012. P. 1.
6. Yeh J.-H., Wu M.-L. // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. V. 1. IEEE Computer Society. 2010. P. 209.
7. Милкова М.А. // Цифровая экономика. 2021. Т. 14. № 2. С. 31.
8. Указ Президента Российской Федерации от 10.10.2019 г. № 490; Официальный интернет-портал правовой информации — <http://www.kremlin.ru/>
9. Blei D.M. // Commun. ACM. 2012. V. 55. № 4. P. 77.
10. Воронцов К.В., Потапенко А.А. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции “Диалог” (Бекасово, 4–8 июня 2014 г.) Вып. 13 (20). М.: Изд-во РГГУ, 2014. С. 676.
11. Blei D.M., Ng A.Y., Jordan M.I. // J. Mach. Learn. Res. 2003. V. 3. P. 993.
12. Краснов Ф.В., Баскакова Е.Н., Смазневич И.С. // Вестник ТГУ. Управление, вычислительная техника и информатика. 2021. № 56. С. 100.
13. ГОСТ Р 56861-2016 “Система управления жизненным циклом. Разработка концепции изделия и технологий”. Москва. 2016.