Научная статья
УДК 004.82: 004.89
https://doi.org/10.24143/2072-9502-2023-2-125-134
EDN VVZRTW

Создание средств интеллектуальной поддержки процедур рекрутинга инновационного предприятия на основе методов обработки естественного языка

Екатерина Алексеевна Машина

Национальный исследовательский университет ИТМО,
Санкт-Петербург, Россия, mashina.katherina@gmail.com

Аннотация. При проведении процедур рекрутинга современной инновационной компании актуальной является задача создания автоматизированных решений, предназначенных для объективного определения компетенций работника. Описаны результаты проведенного аналитического обзора методов, позволяющих выявлять компетенции специалистов на основании семантических исследований порожденных ими текстов, с последующим анализом применимости таких решений для выполнения конкретных процедур рекрутинга. Приведено описание подхода к созданию коллекций текстов, характеризующих компетенции претендента на вакансию, классифицированных по признакам времени генерации текстов (ретроспективный опыт/текущая деятельность) и взаимодействия с соавторами. В качестве инструментальных средств проведения семантического анализа текстов рассмотрены примеры использования различных методов обработки естественного языка, основанных на векторном представлении текстов, для решения конкретных задач, связанных с оценкой компетенций специалистов.

Ключевые слова: инновационное предприятие, кандидат на вакансию, компетенции сотрудника, обработка текстов, коллекция текстов, встречаемость слов

Благодарности: автор выражает искреннюю благодарность коллективу мега-факультета компьютерных технологий и управления Университета ИТМО за предоставленную возможность практической апробации ряда положений данной работы и обсуждения полученных результатов.

Для цитирования: *Машина Е. А.* Создание средств интеллектуальной поддержки процедур рекрутинга инновационного предприятия на основе методов обработки естественного языка // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2023. № 2. С. 125–134. https://doi.org/10.24143/2072-9502-2023-2-125-134. EDN VVZRTW.

Original article

Creating intellectual support tools for recruiting procedures of innovative enterprise by using Natural language processing methods

Ekaterina A. Mashina

ITMO University,
Saint-Petersburg, Russia, mashina.katherina@gmail.com

Abstract. When conducting recruiting procedures for a modern innovative company, the task of creating automated solutions designed to objectively determine the competencies of an employee is most urgent. There have been described the results of an analytical review of methods that allow identifying the competencies of specialists based on semantic studies of texts generated by them, followed by analysis of the applicability of such solutions to perform specific recruiting procedures. There is given a description of an approach to creating collections of texts characterizing the competencies of the applicant for the vacancy, which are classified according to the time of text generation (retrospective experience /current activity) and interaction with co-authors. There have been considered the examples of using different methods of Natural language processing based on the vector representation of texts for solving specific tasks related to the assessment of the competencies of specialists and the tools for semantic analysis of texts.

Keywords: innovative enterprise, applicant for a vacancy, employee competencies, text processing, text collection, frequency of word occurrence

Acknowledgment: the author expresses his sincere gratitude to the staff of the mega-faculty of Computer Technologies and Management of ITMO University for the opportunity to test a number of provisions of this work in practice and discuss the results.

For citation: Mashina E. A. Creating intellectual support tools for recruiting procedures of innovative enterprise by using Natural language processing methods. *Vestnik of Astrakhan State Technical University. Series: Management, computer science and informatics*. 2023;2:125-134. (In Russ.). https://doi.org/10.24143/2072-9502-2023-2-125-134. EDN VVZRTW.

Введение

Происходящий в настоящее время повсеместный переход к технологиям Индустрии 4.0 в качестве основного конкурентного преимущества, обеспечивающего высокий темп инноваций, предполагает создание роботизированных производств во всех областях сервиса и технологий. Движущей силой инновационного развития по-прежнему остается когнитивная деятельность человека, основанная на полученных знаниях и опыте.

Особенно остро проблема подбора необходимых кадров возникает у так называемых инновационных предприятий, для которых удельная доля от реализации инновационного (т. е. не представленного ранее на рынке) продукта или услуги превышает 70 % от общего размера дохода [1]. Отличительными чертами таких компаний являются небольшой размер, работа в еще только формирующихся областях технологий или сервисов, высокая доля использования собственных ноу-хау в инновационных разработках и мультидисциплинарный характер работ.

Высокий уровень зарплат в таких компаниях приводит к тому, что на каждую объявляемую вакансию откликаются до нескольких сотен соискателей, из которых инновационные компании нанимают обычно не более 1 % претендентов [2]. В связи с этим выбор наилучшей кандидатуры становится достаточно сложной задачей, т. к. для проведения объективного и обоснованного выбора необходимого кандидата у компактного инновационного предприятия попросту не хватает временных ресурсов. При этом и со стороны инновационного предприятия часто оказывается достаточно трудно объективно определить и четко описать конкретный набор требований к кандидату.

На сегодняшний день было предпринято большое количество попыток создания программных средств, позволяющих автоматизировать процесс предварительной обработки комплекта подаваемых соискателем текстовых документов, основываясь на различных методах анализа. Подавляющая часть подобных разработок оказывается мало применима в случае инновационных предприятий, поскольку ориентирована на анализ ограниченных по номенклатуре текстов или основывается лишь на анализе формальных признаков образовательной подготовки или производственной квалификации соискателей, в связи с чем эти оценки не позволяют в полной мере отразить личные компетенции соискателей в развивающейся области технологий. Однако квалифицированный специалист в большинстве случаев является автором существенного количества текстовых материалов, отражающих его компетенции.

Поэтому актуальной является задача создания решений, предназначенных для определения компетенций работника на основе анализа семантики порождаемых им текстов.

В связи с этим целью настоящего исследования является проведение аналитического обзора возможных методов, позволяющих выявлять компетенции специалистов на основании изучения порожденных ими текстов, с последующим анализом применимости таких решений для выполнения конкретных процедур рекрутинга.

Подходы к формализации описания компетентностных качеств работника

Вопрос количественного учета и непосредственного использования знаний работников при построении систем управления бизнесом интересовал исследователей уже достаточно давно. С середины прошлого века разрабатывались модели, позволяющие связать роль опыта и знаний работников с темпами роста производства инновационной продукции, в состав которой тем или иным способом входит существенная часть формализованных знаний компании. Так, в работе [3] была описана модель экономического развития, основывающегося на знаниях и инновациях. Позднее для учета производственных организационно-технических улучшений, создаваемых в процессе производства продукции, была предложена модель, описывающая инновационный прогресс в качестве основного способа повышения производительности труда, связанного с накоплением сотрудниками компании добавочного опыта, обусловленного выполнением ими своих непосредственных производственных обязанностей на рабочих местах [4].

Следующим шагом в уточнении описания влияния человеческого капитала на повышение эффективности бизнеса стала модель [5], учитывающая оценку сопоставимости влияния человеческо-

го и материального капитала на процесс производства инноваций и предполагающая, что в капитале предприятия помимо материальных активов существенна роль интеллектуального вклада работников, в сущности представляющего собой овеществление человеческих знаний, существующих во всех промежуточных продуктах, использовавшихся для выпуска данного конечного продукта.

Однако основной прорыв в формальном описании производительных свойств работников предприятия стал возможен лишь в рамках создания концепции единых систем управления корпоративными знаниями, позволившей установить однозначную связь между корпоративными знаниями и их представлением в виде продукта информационных трансформаций, описываемым, в частности, в виде «модели DIMKC» [6] (рис. 1), представляющей собой иерархическую схему информационных переделов, в которой каждый последующий уровень добавляет определенные семантические свойства к предыдущему уровню.

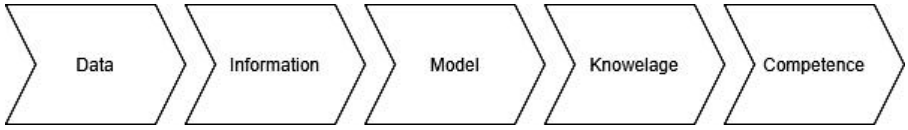

Рис. 1. Методологическая схема информационных трансформаций (информационных переделов), возникающих в процессе управления знаниями (модель DIMKC)

Fig. 1. Methodological scheme of information transformations (information redistribution) arising in the process of knowledge management (DIMKC model)

На основании этого представления Л. Прусаком и Т. Давенпортом [7] была осуществлена одна из первых успешных попыток формального количественного описания элементарных компетенций работника. Авторы [7] предложили количественно учитывать возможность конкретного сотрудника выполнять ту или иную работу как постоянно расширяющуюся в процессе производственной деятельности комбинацию личных образовательных знаний, практического опыта, окружающей контекстной информации, индивидуальных ценностей и интуиции, выражающуюся в том числе в умении вербально описывать проводимые действия.

Это позволило рассматривать текстово-фиксируемые (т. е. проявляющиеся в текстах) индивидуальные компетенции каждого из сотрудников компании как составную величину:

$$K = K0 + K1 + K2\,, \qquad (1)$$

где $K0$ – образовательные компетенции работника; $K1$ – компетенции работника, полученные им в результате предшествовавшей производственной деятельности; $K2$ – соавторские компетенции работника, представляющие собой компетенции его социального (соавторского) окружения, которые он активно использует в своей работе.

Предпосылкой к дальнейшему созданию методики выявления индивидуальных компетенций работника является понимание вербальности человеческого знания и возможности его фиксации в различного рода текстах, созданных средствами естественного языка [8]. Под текстом здесь и далее понимается письменная (или записанная устная) речь, которая внутренне организована и относительно закончена [9].

Чтобы понимать смысл текста, сотрудник предприятия должен быть компетентен в области используемых в рассматриваемом материале понятий, о чем могут свидетельствовать тексты, сгенерированные им ранее.

Таким образом, решение поставленной задачи определения конкретного набора компетенций сотрудника может быть основано на предположении о том, что и компетенции высококвалифицированного работника, и требования вакансии могут быть описаны в виде определенных наборов текстов, на основании сравнительного анализа которых можно сделать вывод о соответствии специалиста его будущему рабочему месту. При определении различных компонент компетенций сотрудника различаться будут лишь в типы подлежащих анализу коллекций текстов (табл.).

Сравнительная оценка отдельных составляющих компетенций сотрудника

Comparative assessment of individual components of an employee's competencies

Объект анализа	**Компетенции, связанные с образованием**	**Компетенции, связанные с опытом работы**	**Компетенции, связанные с соавторством**
Коллекции текстов, подлежащих анализу	Коллекции текстов, порожденных сотрудником в периоды прохождения обучения	Коллекции тестов, описывающие результаты работ, ранее выполненных сотрудником	Коллекции текстов работ, упомянутых в работах сотрудника

Таким образом, представление (1) позволяет свести решение задачи оценки степени соответствия компетенций кандидата на вакансию и требований его будущего рабочего места к решению набора задач по определению степени похожести суммарной коллекции порождаемых работником текстов и коллекций текстов, описывающих его будущее рабочее место.

В первом приближении составные части индивидуальной компетенции соискателя, выявляемые путем анализа порождаемых текстов, рассматриваются как равновесомые. При дальнейшем уточнении описательной модели представления (1) отдельному изучению должны быть подвергнуты процессы редукции компетенций работника [10, 11], связанные как с временем, прошедшим после генерации им конкретного текста (особо характерные для компетенций, связанных с образованием и опытом работы), так и связанные с соавторством. Указанное влияние редукции может быть учтено введением соответствующих коэффициентов при включении конкретных текстов в результирующую коллекцию.

Использование векторного представления текстов для анализа семантической близости коллекций, характеризующих кандидата на вакансию и требования рабочего места

Методологической основой проведения дальнейшего анализа является понимание того, что тексты, сгенерированные на естественном языке, предполагают разную (но определенную) встречаемость слов.

Для проведения подобных процедур сравнения используется векторная модель представления текста (Vector Space Model, VSM). Генерация VSM производится путем представления каждого слова из текста, подвергаемого анализу, в виде многомерного вектора, элементами которого являются слова, используемые во всей выборке.

Существенным является предположение о том, что на содержание текста оказывает влияние лишь частотность использования в нем тех или иных слов, а не места их расположения в тексте.

В этой связи коллекцию подвергаемых анализу текстов можно представить как выборку пар «текст – слово» (d, w), где $d \in D$, $w \in W$, D – множество текстов коллекции, W – множество уникальных терминов этой коллекции (именуемое также используемым словарем).

Одним из простейших способов генерации VSM является «one-hot»-кодировка. Чтобы определить с ее помощью важность конкретного слова в анализируемом тексте, возможно подсчитать, сколько раз унифицированная форма этого слова встречается в тексте, на основании чего определить его смысловую важность, учитывая тот факт, что чем чаще слово встречается в конкретном тексте, тем выше его важность для него (рис. 2).

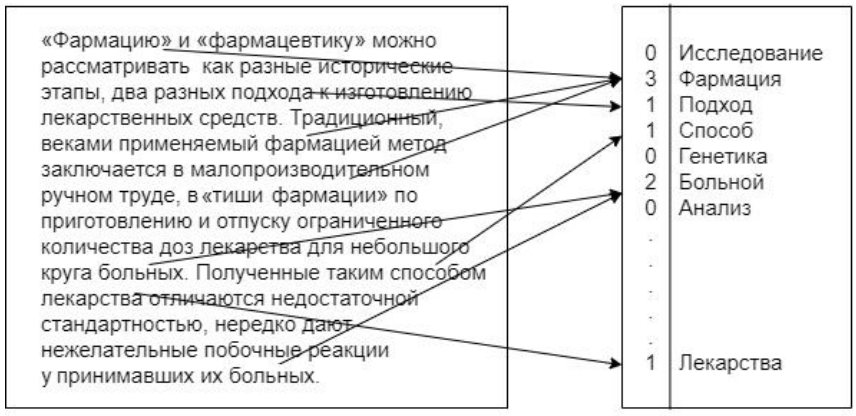

Рис. 2. Определение частотного распределения слов в тексте

Fig. 2. Determining the frequency of word distribution in the text

После того, как у рассматриваемого текста определены веса всех исследуемых слов (в том числе не встречающихся в тексте), возможно построить многомерный вектор, представляющий конкретный текст в векторном пространстве:

$$d_j = \left(w_{1j}, w_{2j}, \ldots, w_{nj}\right),$$

где d_j – векторное представление j-го текста; w_{ij} – вес i-го слова в j-м тексте; n – общее количество различных слов во всех текстах коллекции.

При «one-hot»-кодировке вектор, описывающий конкретное слово, состоит из одной единицы, соответствующей положению слова в словаре, и остальных нулей. Очевидно, что такое сильно

разреженное представление неэффективно по памяти и не позволяет сравнивать слова на предмет семантической близости.

Дальнейшим развитием подходов к созданию VSM стало семейство решений, позволяющих создавать низкоразмерное представление каждого слова по набору анализируемых текстов, а также учитывать контекстуальное подобие слов. Среди таких наиболее производительных методов можно назвать Global Vectors (GloVe), созданный в Стэнфордском университете, а также Word2Vec и Bert, разработанные компанией Google [12]. Указанные методы векторного представления не только могут самостоятельно решать задачи приведения слов к их базовой унифицированной форме (лемме), но и способны учитывать их смысловое подобие, а также различать схожие по написанию слова, встречающиеся в разных контекстах; общепринятые сокращения и аббревиатуры являются для таких представлений объектами, равнозначными словам, формирующим текст. Таким образом, использование высокоуровневых векторных представлений позволит решать широкий круг прикладных задач обработки текстов.

В связи с тем, что и анализируемая коллекция текстов, созданных соискателем, и коллекция текстов, описывающих требования его будущего рабочего места, как правило, характеризуются не слишком большим объемом исходных текстовых данных, при построении методики сравнения для получения векторных представлений возможно использовать языковую модель Global Vectors for Word Representation (GloVe) [13] (далее M).

Так как при проведении анализа используются неразмеченные данные, алгоритм GloVe эффективно применяет статистические особенности совпадений отдельных смысловых элементов текста, минимизируя разницу между логарифмом вероятности совместного появления слов и произведением их векторов, используя метод стохастического градиентного спуска. Это позволяет учесть совместную контекстную встречаемость слов в анализируемых текстах и сгруппировать вектора слов по глобальной схожести.

Используя описанные механизмы, подвергнем процедуре сравнения коллекцию научных текстов A, представленных соискателем вакансии, и коллекцию R, состоящую из текстов, содержащих описание исследовательского проекта, фактически представляющих собой набор квалификационных требований к кандидату.

Возможную близость компетенций кандидата и требований вакансии вычислим через определение косинуса угла между фактическими направлениями многомерных векторов a и r, сгенерированных на основании коллекций A и R, учитывая тот факт, что чем меньше угол между рассматриваемыми векторами, тем выше смысловая близость сравниваемых коллекций текстов.

После обработки коллекции текстов A на языковой модели M определим направление семантического вектора для коллекции текстов соискателя вакансии:

$$a = M(R), \ \forall a \in A,$$

где a – искомый вектор коллекции текстов соискателя; $M(A)$ – обработанная при помощи предобученной модели GloVe коллекция текстов соискателя.

Следующим шагом определим подобный вектор для профессиональных требований к кандидату на открывающуюся вакансию:

$$r = M(R), \forall r \in R,$$

где r – искомый вектор коллекции текстов, описывающих требования вакансии; $M(R)$ – обработанная при помощи предобученной модели GloVe коллекция текстов, описывающих компетентностные требования вакансии.

После этих процедур вычислим интересующую нас косинусную меру смысловой близости $\mathrm{sim}(a, r)$ для векторов a и r, формально характеризующую степень близости компетенций соискателя и компетентностых требований вакансии:

$$\mathrm{sim}(a, r) = \frac{a \cdot r}{\| a \| \| r \|} = \frac{\sum_{i=1}^{n} a_i r_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} r_i^2}}, \quad (2)$$

где a_i, r_i являются компонентами векторов A и R.

Близость косинусной меры к единице будет говорить о близости сравниваемых векторов, а значит, и свидетельствовать о степени соответствий компетенций соискателя требованиям вакансии проекта. Очевидно, что при подобной оценке нескольких соискателей вакансии предпочтение должно быть отдано тому, у которого значение косинусной меры, вычисленной по выражению (2), выше.

Однако при производстве подобного сравнения следует учитывать тот факт, что полученный интегральный результат плохо поддается интерпретированию полученного расчетного показателя, поскольку специалисту hr-подразделения остается непонятным, на основании присутствия каких компетенций предпочтение в конце концов было отдано тому или иному кандидату. Поэтому применение подобной процедуры предпочтительно для построения программных решений, осуществляющих сравнение интегральных компетенций работников «массовых специальностей», а также молодых специалистов.

Повысить степень интерпретируемости получаемого результата сравнения возможно, применяя подходы, основанные на тематическом моделировании текстов, которые позволяют не только осу-

ществить подобное сравнение, но и обеспечить приемлемую для принятия управленческих решений тематическую интерпретируемость результата.

Применение методов тематического моделирования для анализа семантической близости коллекций текстов, характеризующих компетенции кандидата на вакансию

Основной идеей тематического моделирования текстов является представление о том, что наличие каждого конкретного термина в рассматриваемом тексте автора обусловлено необходимостью детального описания в тексте некоторой смысловой темы, неотъемлемой частью контента которой является рассматриваемый термин.

Как и в описанном ранее случае семантического анализа, любой текст автора представляется в виде некоторого неупорядоченного множества слов, для которого существенным является лишь факт наличия слова в исследуемом тексте, а не его конкретное место в лингвистической конструкции. Процесс анализа текста с применением средств тематического моделирования текстов строится на основании предположений о том, что вхождение конкретного слова w в текст d однозначно обосновано некоторой тематикой t из заданного множества тематик T и не зависит от самого текста d, а определяется лишь его тематикой, что может быть описано единым распределением вероятности

$$p(w \mid t) = p(w \mid t, d),$$

где p – распределение вероятности.

В такой постановке исследуемые коллекции текстов (характеризующих научные разработки высококвалифицированного специалиста и требования его будущего рабочего места) могут быть рассмотрены как выборки троек (w_i, d_i, t_i), $i = 1, \ldots, n$ из дискретного распределения $p(w, d, t)$ на конечном множестве декартова произведения $W \times D \times T$, элементами которого являются все возможные упорядоченные тройки исходных элементов.

В каждой из рассматриваемых троек слова w_i и тексты d_i являются наблюдаемыми переменными, а конкретные темы из общего множества тем $t_i \in T$ являются скрытыми (латентными) переменными, которые необходимо определить.

Каждый текст в рассматриваемом случае может быть представлен как некоторое дискретное распределение на множестве тем $\theta_{td} = p(t|d)$, а каждая латентная тема – как дискретное распределение на множестве слов $\varphi_{wt} = p(w|d)$ (рис. 3).

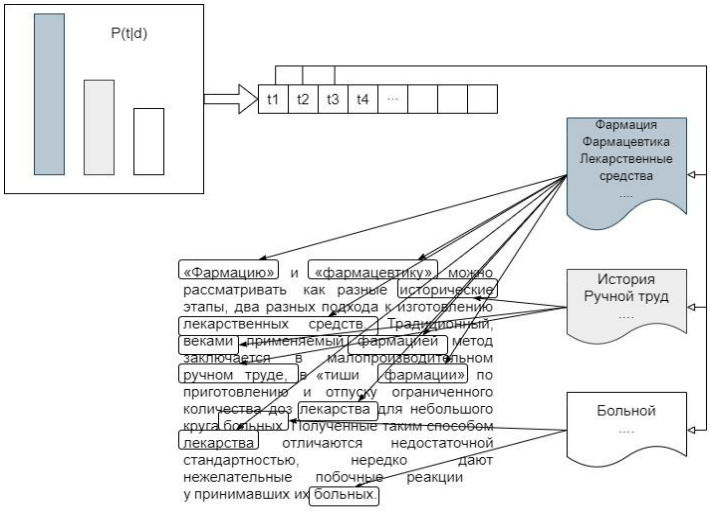

Рис. 3. Схема работы тематического моделирования текста

Fig. 3. Graph of work of thematic text modeling

В этом случае, по теореме о полной вероятности [14], окажется справедливым следующее:

$$p(w \mid d) = \sum_{t \in T} p(w \mid d, t)\, p(t \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d) = \sum_{t \in T} \varphi_{wt}\, \theta td. \qquad (3)$$

С учетом (3) задача построения тематической модели сводится к определению $p(w \mid t)$ для всех $t \in T$ и $p(t \mid d)$ для всех $d \in D$.

После этого, воспользовавшись предположением о том, что распределение слов в конкретной коллекции связано только с темами, а не с текста-

ми, определим логарифм функционала правдоподобия для вероятности совместного появления текста d и слова w в анализируемой коллекции D:

$$L = \sum_{d\in D}\sum_{w\in W} n_{dw} \ln p(d,w) = \sum_{d\in D}\sum_{w\in W} \ln \sum_{t\in D} p(w \mid t)\, p(t \mid d)\, p(d) \to \max p(w \mid t), p(t \mid d) \quad (4)$$

при наличии естественных ограничений

$$\sum_{t\in T} p(t) = 1;\ \sum_{d\in D} p(t \mid d) = 1;\ \sum_{w\in W} p(t \mid w) = 1.$$

Для решения задачи (4) используем, в частности, эффективный итерационный двушаговый алгоритм EM (Expectation-maximization algorithm), широко применяемый при решении задач математический статистики для определения оценок максимального правдоподобия параметров, при наличии у анализируемой модели предполагаемой зависимости от ряда скрытых переменных [15]. Следует иметь в виду, что методы тематического моделирования также самостоятельно проводят предварительную подготовку текстов, связанную с леммированием.

В настоящее время разработано большое количество разнообразных алгоритмов создания решений по анализу больших объемов текстовых коллекций с использованием методик тематического моделирования, которые имеют определенные особенности применения: латентно-семантический анализ (Latent Semantic Analysis, LSA) [16], латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) [17], вероятностный латентно-семантический анализ (Probabilistic Latent Semantic Analysis, PLSA) [18], темпоральные тематические модели (Dynamic topic models) [19] и др.

В настоящее время подобные алгоритмы успешно используются для потокового анализа социальных сетей, в системах выдачи рекомендаций, в большом количестве адаптивных справочных систем. При автоматизации процедур рекрутинга использование методов тематического моделирования представляется особенно предпочтительным при необходимости разработки автоматизированных решений, предназначенных для выявления конкретных компетенций работника в достаточно узкой предметной области.

Применение методов иерархического тематического моделирования анализа семантической близости коллекций текстов, характеризующих компетенции кандидата на вакансию

Признавая высокую результативность тематического анализа профессиональных текстов, характеризующих компетенции высококвалифицированных специалистов, следует отметить, что в условиях управления динамичным и большей частью мультидисциплинарным бизнесом часто требуются механизмы, осуществляющие семантическую обработку текстов на более высоком уровне, позволяющем подробно классифицировать смысловую «подисциплинарную» принадлежность тех или иных тем в конкретных текстах.

Это в первую очередь связано с тем, что при управлении современным инновационным предприятием, включающем формулирование целей, распределение задач, учет возможных рисков, контроль производственных процессов и управление персоналом, фактически параллельно используются два подхода к менеджменту: проектный и функциональный.

Обычно считают, что для инновационных предприятий более всего характерно тяготение к проектному подходу, поскольку деятельность каждого сотрудника такой компании не может быть раз и навсегда четко формализована, и ему приходится работать в условиях постоянных изменений и значительной неопределенности. В связи с этим управленческая структура подобными процессами в большей своей части выстраивается не в виде конкретных процессов или функций, а как некоторый набор работ над конкретной задачей, имеющей даты начала и окончания.

Функциональный подход, основанный на профессиональном принципе распределения труда, где каждый участник производственного процесса четко понимает границы своей ответственности, обеспечивает повышение производительности работ, минимальные издержки на менеджмент, высокую управляемость бизнес-процессов, а следовательно, и высокую эффективность компании.

Поэтому у большинства инновационных компаний наблюдается повышенный интерес к более структурированному (в сторону дифференциации конкретных областей знаний) подходу к определению компетенций сотрудников.

В этой связи становится востребованным построение средств анализа компетенций сотрудников, производимое на основании анализа генерируемых ими текстов, позволяющее более точно кластеризовать монодисциплинарные навыки специалистов (рис. 4), используя для этого методы иерархического тематического моделирования [20].

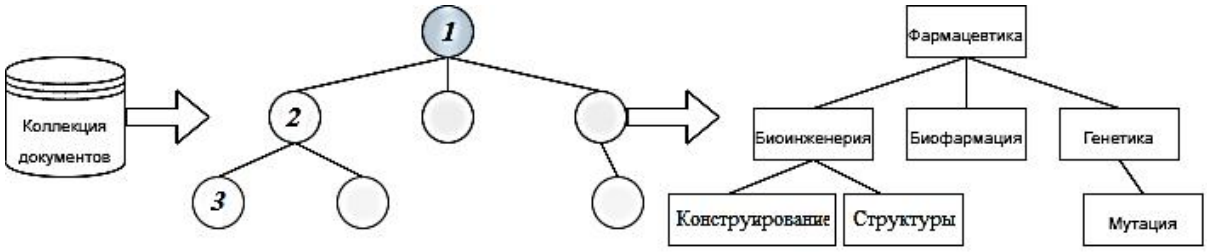

Рис. 4. Построение дисциплинарных тематических иерархий:
1 – корневая вершина; *2* – родительские вершины; *3* – дочерние вершины

Fig. 4. Building the disciplinary thematic hierarchies:
1 – root vertex; *2* – parent vertices; *3* – child vertices

Для построения простейшей иерархии возможно использовать способы, применяемые для плоской кластеризации [21], переходя от понятия «Bag of Words» к понятию «Bag of Topics». Для адекватного построения иерархии тем требуется введение нескольких регуляризаторов, учитывающих ряд дополнительных требований, накладываемых на результирующую модель.

При переходе к нижележащему уровню, который характеризуется множеством содержащихся в нем тем S, генерируется тематическая модель, которая объясняет появление конкретных слов w_i в конкретных текстах d_j, входящих в анализируемую коллекцию с помощью разложения

$$p(w \mid d) = \sum_{s \in S} \sum_{t \in T} p(w \mid s) p(s \mid t) p(t \mid d),$$

где распределение $p(t|d)$ известно, поскольку родительский уровень при переходе к дочернему оказывается уже построенным.

В качестве производительных методов иерархического тематического моделирования для анализа коллекций текстов соискателей могут быть использованы hLDA – hierarchicial Latent Dirichlet Allocation (иерархический LDA) [22], HDP – Hierarchical Dirichlet Processes [20], splitLDA [23], STROD – Scalable Recursive Orthogonal Decomposition [24]. Указанные методы иерархического тематического моделирования оказываются наиболее полезными при создании автоматизированных процедур отбора кандидатов на выполнение инновационных работ смешанной дисциплинарности.

Заключение

В результате проведенного аналитического обзора можно сделать следующие выводы, существенные для проведения дальнейших работ по созданию практических автоматизированных методов оценки компетенций сотрудников предприятий:

– эффективным подходом к анализу компетенций специалиста является совокупность действий, основанная на изучении методами семантического анализа текстово-фиксируемых компетенций, подразделяемых на три составляющие: полученные в процессе профессионального образования, в результате производственного опыта и соавторской активности;

– указанные компетенции специалиста могут быть формально описаны через набор речевых концептов, содержащихся в коллекциях текстов, порождаемых работником, и коллекции текстов, с которыми он взаимодействовал ранее.

Список источников

1. Запорожец А. С. Инновационные предприятия и их особенности с позиций экономической науки // Инновации и инвестиции. 2020. № 10. С. 3–6.

2. Коркина Т. А., Зотова Е. Н. Зарубежный и отечественный опыт подбора персонала // Общество, экономика, управление. 2021. Т. 6, № 4. С. 58–63. DOI: 10.47475/2618-9852-2021-16408.

3. Solow R. M. A contribution to the Theory of Economic Growth // Quarterly Journal of Economic. 1956. Iss. 70. P. 65–94.

4. Arrow K. J. The Economic Implications of Learning by Doing // Review of Economic Studies. 1962. V. 29. P. 155–173.

5. Lucas R. E. On the Mechanics of Economic Development // Journal of Monetary Economics. 1988. 22, July. P. 3–42.

6. Цветков В. Я. Информационное управление. KG, Saarbrücken, Germany: LAP LAMBERT Academic Publishing GmbH&Co, 2012. 201 p.

7. Davenport T., Prusak L. Working Knowledge // Harvard Business Review Press, 2000, 240 p.

8. Mashina E. Uniform Assessment of The Company's Employee's Competence Using Natural Language Processing Methods for Their Further Use in Corporate Knowledge Management Systems // Proceedings of FRUCT'32 – 2022. V. 2. P. 374–381.

9. Гальперин И. Р. Текст как объект лингвистического исследования. М.: КомКнига, 2007. 144 с.

10. Дружинин В. Н. Психология общих способностей. СПб.: Питер, 2007. 368 с.

11. Кутергина Е. А., Санина А. Г. Компетентностные профили чиновников в современной России // Журнал исследований социальной политики. 2017. № 15 (1). С. 113–128.

12. Понкин Д. И. Концепт предобученных языковых моделей в контексте инженерии знаний // International Journal of Open Information Technologies. 2020. V. 8, no. 9. P. 18–27.

13. Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, 2018. P. 328–339.

14. Долгов А. И. Корректные модификации формулы Байеса для параллельного программирования // Суперкомпьютерные технологии: материалы 3-й Всерос. науч.-техн. конф. Ростов н/Д., 2014. Т. 1. С. 122–126.

15. Neal R. M., Hinton G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants // Learning in Graphical Models. Cambridge, MA: MIT Press, 1999. P. 355–368.

16. Dirvester S., Dumay S., Fernas D., Landauer T., Harshman R. Indexing using hidden semantic analysis // Journal of the American Society of Information Sciences. 1990. V. 41 (6). P. 391–407.

17. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. V. 3, no. 4, 5. P. 993–1022.

18. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Тр. Ин-та систем. программирования РАН. 2012. С. 215–242.

19. Blei D. M, Lafferty J. D. Dynamic topic models // Proceedings of the 23rd International Conference on Machine learning ICML'06, 2006. P. 113–120. DOI: 10.1145/1143844.1143859.

20. Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research. 2011. V. 12. P. 12749–2775.

21. Адуенко А., Кузьмин А., Стрижов В. Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Изв. Тульс. гос. ун-та. Естественные науки. 2012. Вып. 4. С. 119–131.

22. Blei D. M., Jordan M., Tenenbaum J. Hierarchical Topic Models and the Nested Chinese Restaurant Process. NIPS, 2003. 8 p. URL: https://www.cs.columbia.edu/~blei/papers/BleiGriffithsJordanTenenbaum2003.pdf (дата обращения: 14.01.2023).

23. Pujara J., Skomorch P. Large-Scale Hierarchical Topic Models // NIPS Workshop on Big Learning. 2012. P. 826–839.

24. Wang C., Liu X., Song Y., Han J. Towards Interactive Construction of Topical Hierarchy: A Recursive Tensor Decomposition Approach // Proc. 2015 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'15), August 2015. P. 1225–1234.

References

1. Zaporozhets A. S. Innovatsionnye predpriiatiia i ikh osobennosti s pozitsii ekonomicheskoi nauki [Innovative enterprises and their features from standpoint of economic science]. *Innovatsii i investitsii*, 2020, no. 10, pp. 3-6.

2. Korkina T. A., Zotova E. N. Zarubezhnyi i otechestvennyi opyt podbora personala [Foreign and domestic experience in recruitment]. *Obshchestvo, ekonomika, upravlenie*, 2021, vol. 6, no. 4, pp. 58-63. DOI: 10.47475/2618-9852-2021-16408.

3. Solow R. M. A contribution to the Theory of Economic Growth. *Quarterly Journal of Economic*, 1956, iss. 70, pp. 65-94.

4. Arrow K. J. The Economic Implications of Learning by Doing. *Review of Economic Studies*, 1962, vol. 29, pp. 155-173.

5. Lucas R. E. On the Mechanics of Economic Development. *Journal of Monetary Economics*, 1988, 22, July, pp. 3-42.

6. Tsvetkov V. Ia. *Informatsionnoe upravlenie* [Information control]. KG, Saarbrücken, Germany: LAP LAMBERT Academic Publishing GmbH&Co, 2012. 201 p.

7. Davenport T., Prusak L. *Working Knowledge*. Harvard Business Review Press, 2000, 240 p.

8. Mashina E. Uniform Assessment of The Company's Employee's Competence Using Natural Language Processing Methods for Their Further Use in Corporate Knowledge Management Systems. *Proceedings of FRUCT'32 – 2022*. Vol. 2. Pp. 374-381.

9. Gal'perin I. R. *Tekst kak ob"ekt lingvisticheskogo issledovaniia* [Text as object of linguistic research]. Moscow, KomKniga Publ., 2007. 144 p.

10. Druzhinin V. N. *Psikhologiia obshchikh sposobnostei* [Psychology of general abilities]. Saint-Petersburg, Piter Publ., 2007. 368 p.

11. Kutergina E. A., Sanina A. G. Kompetentnostnye profili chinovnikov v sovremennoi Rossii [Competence profiles of officials in modern Russia]. *Zhurnal issledovanii sotsial'noi politiki*, 2017, no. 15 (1), pp. 113-128.

12. Ponkin D. I. Kontsept predobuchennykh iazykovykh modelei v kontekste inzhenerii znanii [Concept of pretrained language models in knowledge engineering]. *International Journal of Open Information Technologies*, 2020, vol. 8, no. 9, pp. 18-27.

13. Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Melbourne, 2018. Pp. 328-339.

14. Dolgov A. I. Korrektnye modifikatsii formuly Baiesa dlia parallel'nogo programmirovaniia [Correct modifications of Bayes formula for parallel programming]. *Superkomp'iuternye tekhnologii: materialy 3-i Vserossiiskoi nauchno-tekhnicheskoi konferentsii.* Rostov-on-Don, 2014. Vol. 1. Pp. 122-126.

15. Neal R. M., Hinton G. E. *A view of the EM algorithm that justifies incremental, sparse, and other variants. Learn-*

ing in Graphical Models. Cambridge, MA, MIT Press, 1999. Pp. 355-368.

16. Dirvester S., Dumay S., Fernas D., Landauer T., Harshman R. Indexing using hidden semantic analysis. *Journal of the American Society of Information Sciences*, 1990, vol. 41 (6), pp. 391-407.

17. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, vol. 3, no. 4, 5, pp. 993-1022.

18. Korshunov A., Gomzin A. Tematicheskoe modelirovanie tekstov na estestvennom iazyke [Topic modeling of Natural language texts]. *Trudy Instituta sistemnogo programmirovaniia RAN*, 2012, pp. 215-242.

19. Blei D. M, Lafferty J. D. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine learning ICML'06*, 2006. Pp. 113-120. DOI: 10.1145/1143844.1143859.

20. Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. *Journal of Machine Learning Research*, 2011, vol. 12, pp. 12749-2775.

21. Aduenko A., Kuz'min A., Strizhov V. Vybor priznakov i optimizatsiia metriki pri klasterizatsii kollektsii dokumentov [Feature selection and metric optimization when clustering collection of documents]. *Izvestiia Tul'skogo gosudarstvennogo universiteta. Estestvennye nauki*, 2012, iss. 4, pp. 119-131.

22. Blei D. M., Jordan M., Tenenbaum J. Hierarchical Topic Models and the Nested Chinese Restaurant Process. *NIPS*, 2003. 8 p. Available at: https://www.cs.columbia.edu/~blei/papers/BleiGriffithsJordanTenenbaum2003.pdf (accessed: 14.01.2023).

23. Pujara J., Skomorch P. Large-Scale Hierarchical Topic Models. *NIPS Workshop on Big Learning*, 2012, pp. 826-839.

24. Wang C., Liu X., Song Y., Han J. Towards Interactive Construction of Topical Hierarchy: A Recursive Tensor Decomposition Approach. *Proc. 2015 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'15), August 2015.* Pp. 1225-1234.

Статья поступила в редакцию 27.02.2023; одобрена после рецензирования 05.04.2023; принята к публикации 21.04.2023
The article is submitted 27.02.2023; approved after reviewing 05.04.2023; accepted for publication 21.04.2023

Информация об авторе / Information about the author

Екатерина Алексеевна Машина – преподаватель факультета программной инженерии и компьютерной техники; Национальный исследовательский университет ИТМО; mashina.katherina@gmail.com

Ekaterina A. Mashina – Lecturer of the Faculty of Software Engineering and Computer Technology; ITMO University; mashina.katherina@gmail.com