УДК 004.891:004.9

МОДЕЛЬ ОПРЕДЕЛЕНИЯ НОВИЗНЫ ЗАЯВКИ НА ПРОЕКТНОЕ ФИНАНСИРОВАНИЕ МЕТОДАМИ ТЕРМИНОЛОГИЧЕСКОГО АНАЛИЗА

Сироткин А.В., Копченко В.К.

ФГБОУ ВО «Северо-Восточный государственный университет», Магадан, e-mail: andrew_sirotkin@mail.ru, vkkopchenko49@gmail.com

В работе описана концепция разработки автоматизированной системы распределения грантов (АСРГ), разработана процессно-поточная модель IFAR (*Idea, Formalization, Analysis of the Request*) тематического анализа заявки на основе терминологической экспертизы для решения задач отраслевой идентификации проекта, анализа заимствований и оценки новизны, описаны этапы принятия решения о допуске заявки к экспертной оценке. Определение отраслевой принадлежности заявки реализовано с помощью измененной метрики TF-IDF, основанной на поиске взвешенных термов в тексте документа. Для анализа проекта на уникальность авторами было принято решение использовать алгоритм шинглов, работа которого основана на составлении последовательностей слов нормализованного текста и расчета процента неодинаковых последовательностей от общего их количества в проверяемой заявке и эталонном тексте. Для решения задачи определения новизны разработаны соответствующие математические модели. Описана формальная модель заявки, основанная на структуре формулы изобретения и состоящая из детерминированных атрибутов, упрощающих извлечение требуемой информации о проекте. Даны определения новизны и отличительного признака. Рассмотрен способ определения новизны проекта, основанный на кластеризации с помощью самоорганизующихся карт Кохонена, дано обоснование отказа от использования данного метода. Рассмотрены способы определения отличительных особенностей заявки, основанные на методах релевантного поиска. Разработаны критерии наличия отличий и наличия новизны, реализованные на основе терминологического анализа. Описаны некоторые особенности программной реализации АСРГ.

Ключевые слова: грант, заявка, новизна, терминологический анализ, терм, вес, семантика

MODEL FOR DETERMINING THE NOVELTY OF A REQUEST FOR PROJECT FINANCING VIA TERMINOLOGICAL ANALYSIS METHODS

Sirotkin A.V., Kopchenko V.K.

North-Eastern State University, Magadan, e-mail: andrew_sirotkin@mail.ru, vkkopchenko49@gmail.com

The article describes the concept of an automated grant distribution system (ASDG) development. The process-flow model *IFAR* (*Idea, Formalization, Analysis of the Request*) for thematic analysis of the application based on terminological expertise, was described. This model was developed to solve the problems of industry identification of the project, analysis of borrowings and novelty assessment. The stages of deciding on the admission of the application to expert evaluation were described. The industry affiliation of the application was implemented via the modified *TF-IDF* metric, based on the search for weighted terms in the document text. To analyze the project for uniqueness, the authors decided to use the shingle algorithm, which is based on compiling sequences of words in a normalized text and calculating the percentage of different sequences from the total number of them in the checked application and the reference text. To solve the problem of determining the novelty, a mathematical model were developed. A formal application model, based on the structure of the claim and consisting of deterministic attributes that simplify the extraction of the required information about the project, was developed. Definitions of novelty and distinctive feature were given. A method for determining the novelty of the project based on clustering using self-organizing Kohonen maps was considered, and a justification for refusing to use this method was given. The methods of determining the distinctive features of the application based on relevant search methods were considered. The criteria for differences and novelty, implemented based on terminological analysis, and some features of the ASRG software implementation were described.

Keywords: grant, application, novelty, terminological analysis, term, weight, semantics

В 2019 г. Президентом Российской Федерации была утверждена национальная стратегия развития искусственного интеллекта до 2030 г., в перечень задач которой входит разработка интеллектуального программного обеспечения. В рамках реализации этой стратегии представляется актуальной разработка программного обеспечения для проведения экспертных оценок, в частности, как одного из представителей такого класса задач, автоматизированной системы распределения грантов на проектное финансирование.

Попыткой решения этой задачи можно считать создание автоматизированной системы распределения грантов (АСРГ), концепция разработки которой сформулирована авторами в работе [1]. Целью системы является удовлетворение заявки на проектное финансирование на основе экспертного анализа состава заявки, её целей, новизны, ожидаемого эффекта и пр. Авторы данной разработки позволили себе смелость назвать модель проекта Idea, Formalization, Analysis of the Request – IFAR, каковая аббревиатура будет далее использоваться для упоминания о системе.

Целью исследования является разработка автоматизированной информационной

системы распределения грантов. Предметом исследования в рамках данной работы является модель определения новизны заявки на проектное финансирование.

Материалы и методы исследования

Предварительный анализ заявки на грантораспределение, предшествующий, собственно, экспертной оценке, по мнению авторов, включает в себя следующие этапы:

1. Определение отраслевой принадлежности заявки. Постановка этой задачи представлена в работе [2]. Идея такой идентификации исходит из универсального характера системы АСРГ, не ориентированного на какую-либо конкретную область деятельности человека. Поскольку последующие оценки основаны на тематическом анализе, требующем привлечения соответствующих тезаурусов, необходимо заранее определить тематическую область, что значительно сократит время, затрачиваемое на анализ. Эта задача решается на этапе определения отраслевой принадлежности.

2. Анализ оригинальности заявки. Включает в себя определение целей анализа на основе целей заявки, сравнение с другими заявками, накопленными в системе, и при необходимости сравнение с иными документами, определёнными целеполаганием заявки.

3. Анализ новизны предлагаемого решения. Этот этап необходим для установления оригинальности предлагаемого заявителем решения и, по мнению разработчиков, представляет некоторый интерес с точки зрения своей реализации, что и является предметом настоящей работы.

Анализ заявки, реализующий три перечисленных этапа, можно представить в виде процессно-поточной модели тематического анализа, схема которой представлена на рис. 1. При создании модели авторы позволили себе отойти от канонических графических примитивов для представления идеи и использовать свои, наилучшим образом отражающие семантику проекта.

Согласно модели принятие решения о допуске заявки к экспертной оценке состоит из двух этапов:

1. Подготовительный. На данном этапе формируется идея проекта, которая впоследствии формулируется в заявку на финансирование. Текст заявки приводится к каноническому виду, пригодному для дальнейшего тематического анализа, удаляются стоп-символы, стоп-слова, проводится лемматизация [3] каждого термина. Подобная подготовка необходима для дальнейшей обработки заявки методами терминологического анализа.

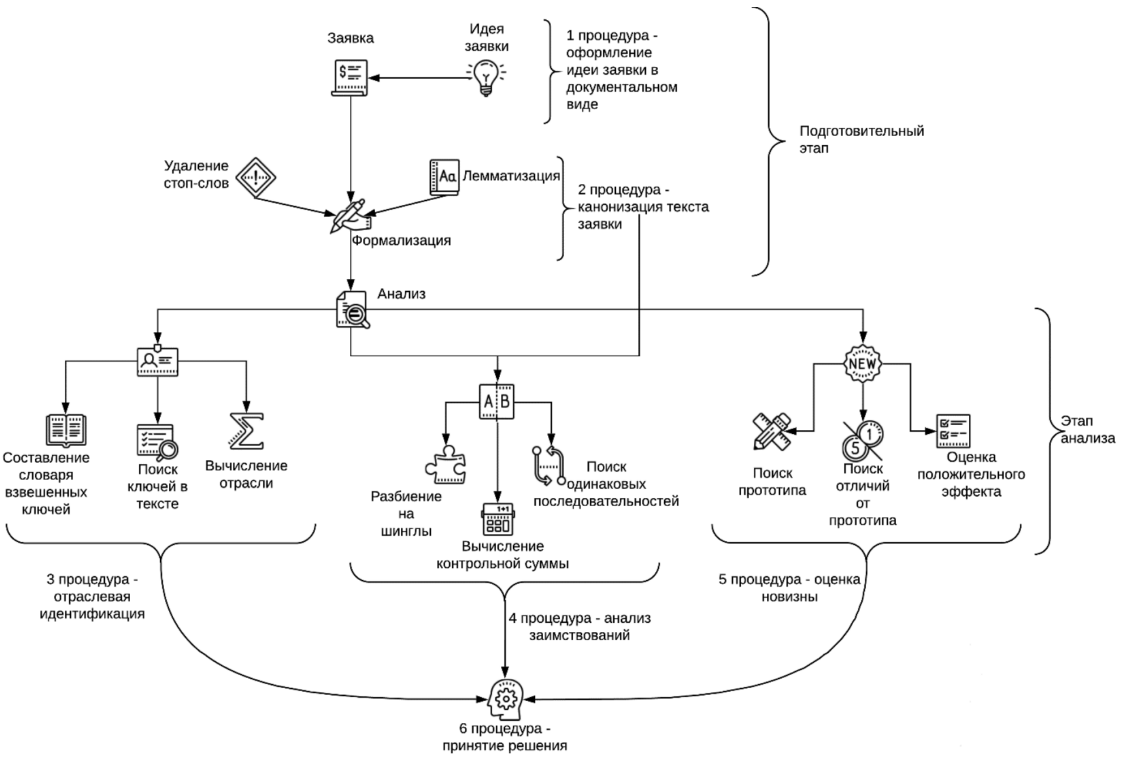

Рис. 1. Модель предварительного анализа заявки

2. Предметный. На данном этапе анализируется канонический текст заявки, определяется отрасль деятельности, в рамках которой разрабатывается проект, оценивается уникальность проекта и его новизна. Расчет значений указанных метрик заявки является необходимым и достаточным для принятия решения о допуске заявки к дальнейшему прохождению экспертной оценки или отказе в таковом.

Для решения поставленных задач отраслевой идентификации, анализа уникальности и оценки новизны требуется разработать модель заявки. Поскольку данный тип документов не обладает очевидными атрибутами, требуемыми для анализа, в первую очередь следует формализовать заявку. Пусть *Docs* – множество заявок на проектное финансирование, хранящихся в системе, *Req* – элементы множества. Коллекцию *Docs* можно представить в виде множества

$$Docs = \{Req_1, Req_2, .., Req_k\}, \quad (1)$$

где $k = |Docs|$.

Модель заявки можно составить на основе формулы изобретения [4]. Этот механизм формализации и определения новизны для заявок на изобретения, прошедший столетнюю апробацию, показал свою высокую эффективность и может быть принят нами как прототип данного решения. Мы назовём данный механизм «формулой заявки».

Состав формулы заявки, сформулированный по методике формулы изобретения, на наш взгляд, наилучшим образом будет отражать признаки – характеристики проекта, участвующего в розыгрыше гранта. В частности, формула заявки предусматривает определение вида проекта и явного перечисления качеств, способных эффективно повлиять на конечный результат, причем для каждого вида заявок определяется свой набор показателей. Заявку можно представить в виде модели

$$Req = <F, C, V>, \quad (2)$$

где F – формула заявки, C – детальное описание проекта, V – атрибуты заявки, определяемые АСРГ в процессе анализа. Формула заявки может быть представлена следующим кортежем:

$$F = <F_T, F_S, F_N>, \quad (3)$$

где F_T – наименование проекта; F_S – краткое описание проекта общими признаками, F_N – признаки, обеспечивающие новизну решения, причем все признаки представлены термами. Атрибуты АСРГ сопровождаются кортежем

$$V = <I, U, H>, \quad (4)$$

где I – отрасль заявки; U – уникальность заявки; H – величина новизны проекта. Отраслевая идентификация определяет величину r соответствия заявки g-й отрасли деятельности и рассчитывается по формуле

$$r = \max \sum_{i=1}^{|G|} w_{ij}, \quad (5)$$

где w_{ij} – вес j-го терма в i-й отрасли, G – множество отраслей, $G = \{g_1, g_2, \ldots, g_c\}$, $j = \overline{1, |g_i|}$. Авторами данный метод был реализован в программном исполнении и показал высокую эффективность.

Анализ заявки на плагиат реализуется с помощью алгоритма шинглов, эффективность которого рассмотрена в работе [5]. Данный алгоритм разбивает канонический текст анализируемой заявки и тексты из архива документов на последовательности, называемые шинглами, причем размер шингла обратно пропорционален эффективности поиска дубликатов. После формирования подстрок вычисляются контрольные суммы последовательностей, совпадение которых между текстом заявки и архивным текстом снижает оригинальность рассматриваемого документа.

Для задачи оценки новизны, несмотря на широкое освещение в научной публицистике (например, [6–8]), не существует универсального решения, так как в большинстве случаев не сформированы наборы показателей для многопараметрических подходов. Например, в работе [9] авторы при решении задачи ранжирования новостей по критерию новизны предлагают учитывать дату публикации новости и скорость появления похожих новостей в сети Интернет. В работе [10] автор предлагает решать задачу оценки новизны через составление словаря маркеров – слов, регулярно появляющихся в авторефератах диссертаций, и дальнейший поиск маркеров в тексте, для которого требуется провести оценку новизны. Широкое распространение получили такие методы выявления новизны, как кластеризация и релевантный анализ. Одним из наиболее известных методов кластеризации является построение самоорганизующейся карты Кохонена (в научной публицистике также встречается обозначение «нейронная сеть Кохонена») [11]. Карта Кохонена распознает схожие элементы в обучающих данных и относит все данные к тем или иным группам, близким по своему содержанию. Обучающие данные для карты формируются на основе векторной модели TF-IDF. Если после проведения кластеризации учебных данных карта обнаружит набор, который по своим характеристикам не принадлежит ни одной из сформированных групп, то она не сможет классифици-

ровать такой набор и тем самым выявит его новизну. Сеть Кохонена работает по принципу соревнования – нейроны второго слоя соревнуются друг с другом за право наилучшим образом сочетаться с входным вектором сигналов [12]. Мерой близости двух векторов чаще всего выступает Евклидова метрика:

$$p(x, y) = \sum_{j=i}^{n} (x_j - y_j)^2. \quad (6)$$

Несмотря на широкое применение кластерного анализа в задачах выявления новизны, данный метод более пригоден для выявления отличий. Не каждое отличие приводит к формированию положительного эффекта или эффективности его достижения, в связи с чем авторами было принято решение отказаться от кластерного анализа. На этом основании была сформулирована задача разработки собственного решения, использующего другие подходы к анализу новизны, которые также нужно было разработать.

Новизна – это совокупность новых качеств, устанавливающих достижение положительного эффекта. Качественная оценка таких качеств позволяет выявить актуальность проекта и подтвердить необходимость его реализации. Таким образом, новизну проекта можно представить в виде

$$H = (D, N) \rightarrow \max, \quad (7)$$

где $N = \{n_i\}$ – множество качеств, выгодно отличающих проект, D – условие отличия анализируемой заявки от множества других (8), которое выполняется при $D > 0$. Отличие – это признак, создающий разницу между объектами, без оценки качества этого отличия.

$$D = dif(x, Y), \quad (8)$$

где $x = \overline{1,|Y|}$; $Y \in O$, где O – отрасль деятельности.

Процесс решения задачи оценки новизны состоит из трех этапов:

1. Поиск прототипа рассматриваемого предмета.

2. Поиск отличий предмета от прототипа.

3. Оценка декларируемого положительного эффекта.

Поиском прототипа можно считать поиск такой заявки *Req* из множества *Docs*, у которой признаки, обеспечивающие новизну проекта, встречаются в других заявках наиболее часто. Множество таких признаков лаконично, элементы в нем не повторяются, что позволяет на первом этапе отказаться от использования метрики *TF-IDF* [13] в силу отсутствия необходимости учитывать частоту терма. Анализировать список можно только по количеству совпадений признаков. Пусть *Sim* – количество совпадений признаков из формулы рассматриваемой *i*-й заявки в списке новшеств заявки из коллекции *Docs*, *d* – прототип анализируемого проекта, тогда совпадения можно рассчитать по формуле

$$d_i = \max Sim_j, \quad (9)$$

где $Sim_j = F_{N_i} \cap F_{N_j}$, $j \in \overline{[1,|Docs|]}$. В случае отсутствия результата можно применить оценку важности признаков. Для этого относительно рассматриваемого множества термов описательной части заявки C сопоставляем каждому признаку (терму) k_i величину w_i, называемую весом. В этом случае прототипом d проекта, претендующего на грант, можно считать проект, удовлетворяющий выражению

$$d = \arg\max \sum_{k=1}^{|Docs|} w_{kj}, \quad (10)$$

где k – терм в детальном описании проекта, w – вес терма, $j \in \overline{[1,|C_i|]}$. Установим критерий оценки новизны – *Criteria*:

$$Criteria = O(F_{N_j}; F_{N_i}), \quad (11)$$

где $O(F_{N_j}; F_{N_i})$ – разность между количеством идентичных признаков новизны в заявке из коллекции и количеством признаков новизны в рассматриваемой заявке.

Установим, что величина совпадений новых признаков в рассматриваемой заявке с признаками, предложенными в архивных документах, не может превышать 50 %. Если результат не удовлетворяет данному условию, то есть $Criteria \geq \frac{|F_{N_i}|}{2}$, то проект не соответствует критериям новизны, грант в таком случае не может быть одобрен.

Данная модель была реализована в виде веб-приложения на базе архитектуры *Application Server* и стека *FAMP – FreeBSD*, *Apache*, *MySQL*, *PHP*. Выбор данных средств обусловлен их популярностью, надежностью, гибкостью в настройке и администрировании [14]. В качестве низкоуровневой подсистемы *MySQL* была выбрана *InnoDB*, так как данная подсистема поддерживает механизмы транзакций, внешних ключей и полнотекстового поиска. Для работы с MySQL используется библиотека объектно-реляционного отображения *RedBeanPHP*, которая автоматизирует процесс создания, редактирования и удаления баз данных, таблиц, записей в зависимости от состава данных, подлежащих сохранению.

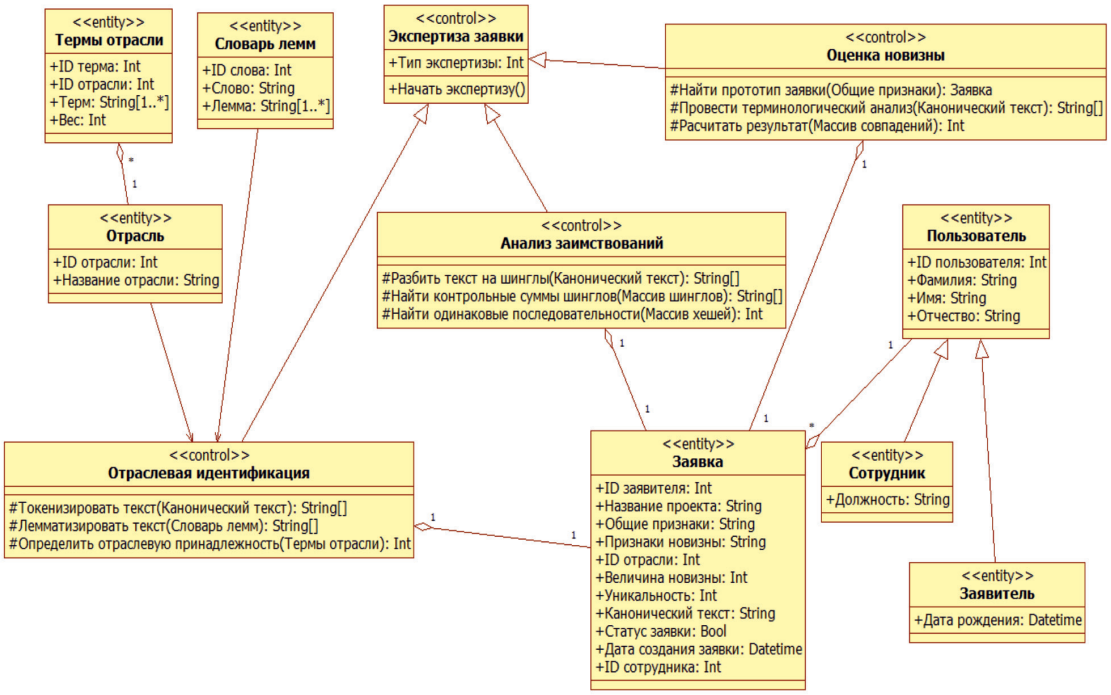

Рис. 2. Взаимодействие компонентов

Схема взаимодействия компонентов системы представлена на рис. 2. В качестве шаблона проектирования был выбран полиморфизм.

В соответствии с данным шаблоном каждый тип экспертизы является производным классом от управляющего класса «Экспертиза заявки» и характеризуется своими операциями. Классы-сущности «Сотрудник» и «Заявитель» являются производными от класса «Пользователь», в котором описаны базовые атрибуты всех акторов системы. Диаграмма также отражает мощность отношений между классами проектирования. Например, каждая отрасль деятельности в рамках АСРГ представлена множеством взвешенных тематических ключей, экспертная оценка значимости которых определяет результаты анализа на отраслевую принадлежность. Заявка принадлежит только одному заявителю и рассматривается одним сотрудником.

Выводы

Разработанные критерии и модель помимо данной задачи также могут быть применены для создания программных средств, ориентированных на оценку новизны документированных предметов, а также могут быть использованы в решении иных задач, нацеленных на выявление качественных отличий. Представленная модель была реализована программными средствами и показала высокую эффективность решения.

Список литературы

1. Сироткин А.В., Копченко В.К. Концепция разработки автоматизированной системы распределения грантов // Eurasiascience: материалы XXI Международной научно-практической конференции (Москва, 15 мая 2019 г.). М.: Актуальность.рф, 2019. С. 107–109.

2. Сироткин А.В., Старикова О.А. Отраслевая идентификация заявок в автоматизированной экспертной системе распределения грантов // Современные наукоёмкие технологии. 2019. № 7. С. 99–103.

3. Жердева М.В., Артюшенко В.М. Стемминг и лемматизация в lucene.net // Вестник МГУЛ – Лесной вестник. 2016. № 3. С. 131–134.

4. Супотницкий М.В. Формула изобретения // Ведомости Научного центра экспертизы средств медицинского применения. 2013. № 1. С. 41–44.

5. Broder A. Identifying and Filtering Near-Duplicate Documents. COM'00: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. Springer-Verlag. 2000. P. 1–10.

6. Sendhilkumar S., Nachiyar N., Mahalakshmi G.S. Novelty Detection via Topic Modeling in Research Articles. Computer Science & Information Technology. 2013. No. 3. P. 401–410. DOI: 10.5121/csit.2013.3542.

7. Salton G., MacGill M.J. Introduction to modern information retrieval. N.Y.: McGraw-Hill, 1983. 448 p.

8. Полонский В.М. Определение новизны результатов научно-педагогических исследований // Проблемы современного образования. 2011. № 2. С. 61–70.

9. Del Corso G.M., Gullí A., Romani F. Ranking a stream of news. International World Wide Web Conference. Proceedings of the 14th international conference on World Wide Web. 2005. P. 97–106.

10. Толчеев В.О. Автоматизированное оценивание формулировок научной новизны публикаций // Заводская лаборатория. Диагностика материалов. 2017. № 83 (5). С. 72–78.

11. Павлова А.И., Синельникова А.С., Чентаева Е.А. Исследование самоорганизующихся карт Кохонена // Современные материалы, техника и технологии. 2015. № 1 (1). С. 184–187.

12. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 400 с.

13. Михайлов Д.В., Козлов А.П., Емельянов Г.М. Выделение знаний и языковых форм их выражения на множестве тематических текстов: подход на основе меры tf-idf // Компьютерная оптика. 2015. № 3. С. 429–438.

14. Колисниченко Д.Н. FreeBSD. От новичка к профессионалу. СПб.: БХВ-Петербург, 2012. 602 с.