УДК 004.891:004.9

ОТРАСЛЕВАЯ ИДЕНТИФИКАЦИЯ ЗАЯВОК В АВТОМАТИЗИРОВАННОЙ ЭКСПЕРТНОЙ СИСТЕМЕ РАСПРЕДЕЛЕНИЯ ГРАНТОВ

Сироткин А.В., Старикова О.А.

Северо-Восточный государственный университет, Магадан, e-mail: andrew_sirotkin@mail.ru, star-olga@yandex.ru

Концептуализируется разработка автоматизированной системы распределения финансовых грантов, основанной на использовании искусственного интеллекта в части проведения экспертных оценок. Роль экспертов по оценке параметров заявки возлагается на автоматизированную систему, результатом работы которой является обобщённый показатель, учитывающий численные параметры соответствия заявки установленным критериям. В качестве предварительной процедуры предлагается использовать определение отраслевой принадлежности заявки, по результатам которой для проведения экспертной оценки будут выбраны критерии и процедуры, специфические для этой области знания. Процедура отраслевой идентификации основана на использовании тематической статистической модели, использующей фиксированные, заранее разработанные наборы ключевых признаков (идентификаторов), дополненные установленными для каждого ключа весами. Разработаны концептуальные модели процесса рассмотрения заявки на грант с помощью автоматизированной системы и математические модели для определения принадлежности заявки к определённой области знания. В качестве таких областей использованы классификаторы ГРНТИ, УДК и др. Степень соответствия заявки области знания устанавливается путем сопоставления множеств признаков, характеризующих области знания, и частоты вхождения их экземпляров в текст заявки. Сравниваются модели тематического анализа, в том числе векторная, и модели, учитывающие содержательное разнообразие и/или усреднённую длину текста. Представлен выбор одной области знания из нескольких, определённых в ходе идентификации.

Ключевые слова: тематический анализ, экспертная оценка, грант, заявка

BRANCH IDENTIFICATION OF DEMANDS IN THE AUTOMATED EXPERT SYSTEM OF DISTRIBUTION OF GRANTS

Sirotkin A.V., Starikova O.A.

North-Eastern State University, Magadan, e-mail: andrew_sirotkin@mail.ru, star-olga@yandex.ru

Conceptualizes working out of the automated system of distribution of the financial grants, based on use of an artificial intellect regarding carrying out of expert estimations. The role of experts according to demand parametres is assigned to the automated system which result of work is the generalised indicator considering numerical parametres of conformity of the demand to installed criteria. As preliminary procedure it is offered to use definition of a branch accessory of the demand by which results for carrying out of an expert estimation criteria and the procedures specific to this area of knowledge will be chosen. Procedure of branch identification is based on use of the thematic statistical model using fixed, in advance developed sets of key signs (identifiers), added with the scales installed for everyone key. By means of the automated system and mathematical models conceptual models of process of consideration of the demand for the grant are developed for definition of an accessory of the demand to certain area of knowledge. As such areas qualifiers GRNTI are used, UDC, etc. Degree of conformity of the demand of area of knowledge is installed by comparison of sets of the signs characterising areas of knowledge, and frequency of occurrence of their copies in the demand text. Models of the thematic analysis, including vector and the models considering a substantial variety and-or average length of the text are compared. The choice of one area of knowledge from several, defined during identification is presented.

Keywords: the thematic analysis, expert estimation, the grant, the demand

В настоящее время возросло количество задач, доверяемых решению автоматизированных экспертных системам. Следует отметить возрастание сложности задач и, как следствие, появление в них интеллектуальных систем, призванных решать задачи, основанные на неформализуемых или плохо формализуемых критериях, требующих использования не только знаний, но и систем, имитирующих работу эксперта, то есть его опыт, интуицию и пр.

Существуют задачи, решение которых требует анализа формализованных критериев наряду с неформализованными, относящихся к узкой области знаний или деятельности человека и общества, оценка принадлежности к которой также требует работы эксперта. К классу таких задач, например, можно отнести рассмотрение и удовлетворение заявок на проектное финансирование по отраслям деятельности, иначе называемое распределение грантов. Анализ и повышение эффективности решения таких задач имеют высокую актуальность, что отражается в работах исследователей, в частности [1].

В подобного вида деятельности практикуется привлечение экспертов, при том, что для этого класса задач не наблюдается широкого освещения в научной публицистике инфор-

мации о разработке и использовании искусственных систем, основанных на автоматизированных технических решениях. В силу этого возникает возможность формулирования задачи построения автоматизированной системы распределения грантов (АСРГ) с использованием системы искусственного интеллекта, работающей на основе разработанных шаблонов по отраслям знаний.

Работа эксперта по рассмотрению заявки на грант достаточно типизирована независимо от предметной области. Как правило, в современных условиях, с использованием *Web*-технологий, процесс удовлетворения заявки и взаимодействия с заявителем подчиняется сценарию, приведённому на рис. 1.

1. Заявитель самостоятельно направляет заявку грантооператору с использованием *Web*-формы.

2. Сотрудник грантооператора проводит первичную оценку заявки на соответствие базовым требованиям.

3. Эксперты проводят анализ заявки на соответствие требованиям и выносят частные решения об удовлетворении заявки.

4. Производится обобщение частных решений и принимается решение об удовлетворении заявки (как правило, сотрудником грантооператора).

5. Заявитель получает уведомление о результатах удовлетворения заявки.

Подобная модель действует для большинства предметных областей, различия могут быть в составе критериев, знаниях, используемых для оценки, а также в составе некоторых процедур оценки, характерных для конкретной отрасли.

Цель исследования: разработка автоматизированной информационной системы распределения грантов. Предметом исследования в рамках данной работы является разработка методики определения принадлежности заявки к конкретной области знания.

Для достижения поставленной цели был проведён анализ процесса рассмотрения заявки в моделях *AS-IS* и концептуализирован процесс автоматизированного рассмотрения в модели *TO-BE*.

Организационно-процессная модель удовлетворения заявки (на грант) Θ в модели *AS-IS*, задающая контекст отношений целевого процесса, может быть представлена следующим образом:

$$\Theta_{AS-IS} = \left\langle \Psi, \Omega_S, P_p, M_\Omega, Q_p, Q_M, O_M \right\rangle, \quad (1)$$

где Ψ – множество заявок, Ω_S – множество экспертов, P_p – множество процедур предварительной оценки (на соответствие требованиям), M_Ω – множество процедур экспертной оценки, Q_p – множество критериев для предварительной оценки (на соответствие требованиям), Q_M – множество критериев для экспертной оценки, O_M – множество численных оценок.

Рис. 1. Диаграмма поточной модели (AS-IS) процесса удовлетворения заявки на грант

Статическая модель эксперта в модели *AS-IS* может быть представлена следующим выражением

$$S_{AS-IS} \Rightarrow (k, \mathrm{Y}, \Phi), \qquad (2)$$

где k – идентификатор эксперта, Y – описание атрибутов эксперта, Ф – значение компетенции эксперта в соответствующей области знаний (фиксированное максимальное значение на текущий момент времени).

Исходя из модели (1–2) можно сделать предположение, что возможно построение универсальной системы распределения грантов, пригодной для использования в любой прикладной области, сопровождаемой кортежем

$$A = \langle \Lambda, M, Q_M, O_M, B \rangle, \qquad (3)$$

где Λ – база знаний и фактов для проведения экспертной оценки по не формализуемым или плохо формализуемым критериям, B – отраслевой шаблон экспертной оценки, который устанавливает процедурные различия для проведения оценки для различных предметных областей. Данная модель, в отличие от модели (1–2), содержит только одну процедуру экспертной оценки, поскольку в системе присутствует только один эксперт – АСРГ. Соответственно (3), организационно-процессная модель удовлетворения заявки (на грант) Θ в модели *TO-BE* может быть представлена следующим образом:

$$\Theta_{TO-BE} = \langle \Psi, P_p, M, Q_p, Q_M, O_M \rangle. \qquad (4)$$

Приняв данную модель, можно сформулировать задачу разработки универсальной АСГР, контекстная диаграмма которой представлена на рис. 2.

В пояснение к диаграмме были сформулированы следующие аргументы.

Основным процессом в данной области деятельности является процесс рассмотрения заявки. Отклонение заявки (её неудовлетворение) не является целью и может рассматриваться как отрицательный результат. Целью грантодателя является именно удовлетворение заявок, соответствующих по наборам критериев требованиям выделения средств. Поэтому следует акцентировать внимание на том, что именно удовлетворение, но не рассмотрение заявки является целью всего процесса.

Статическая модель критериев для проведения экспертной оценки Q_M может быть построена на множествах $|\Omega| = 1$ формализованных $\dot{E}_M$ и неформализованных $\ddot{E}_M$ критериев как

$$Q_M \Rightarrow (\dot{E}_M, \ddot{E}_M).$$

Статическая модель экспертных оценок O_M строится на основе множеств оценок по формализованным критериям $B_{\dot{E}_M}$ и множестве оценок по неформализованным или плохоформализованным критериям, требующим интеллектуальной оценки, $B_{\ddot{E}_M}$

$$O_M \Rightarrow (B_{\dot{E}_M}, B_{\ddot{E}_M}).$$

Модель процесса экспертного оценивания может быть построена как для традиционной модели (1–2), так и для АСРГ (3–4) при условии $|\Omega| = 1$. Её вид может быть представлен следующим кортежем:

$$M_\Omega = \langle \Psi, \dot{E}_M, \ddot{E}_M, R_{\dot{E}_M}, R_{\ddot{E}_M}, B_{\dot{E}_M}, B_{\ddot{E}_M} \rangle, \qquad (5)$$

где R – процесс оценивания на множествах $\dot{E}_M$ или $\ddot{E}_M$.

Рис. 2. Главная контекстная диаграмма «Работа универсальной АСРГ»

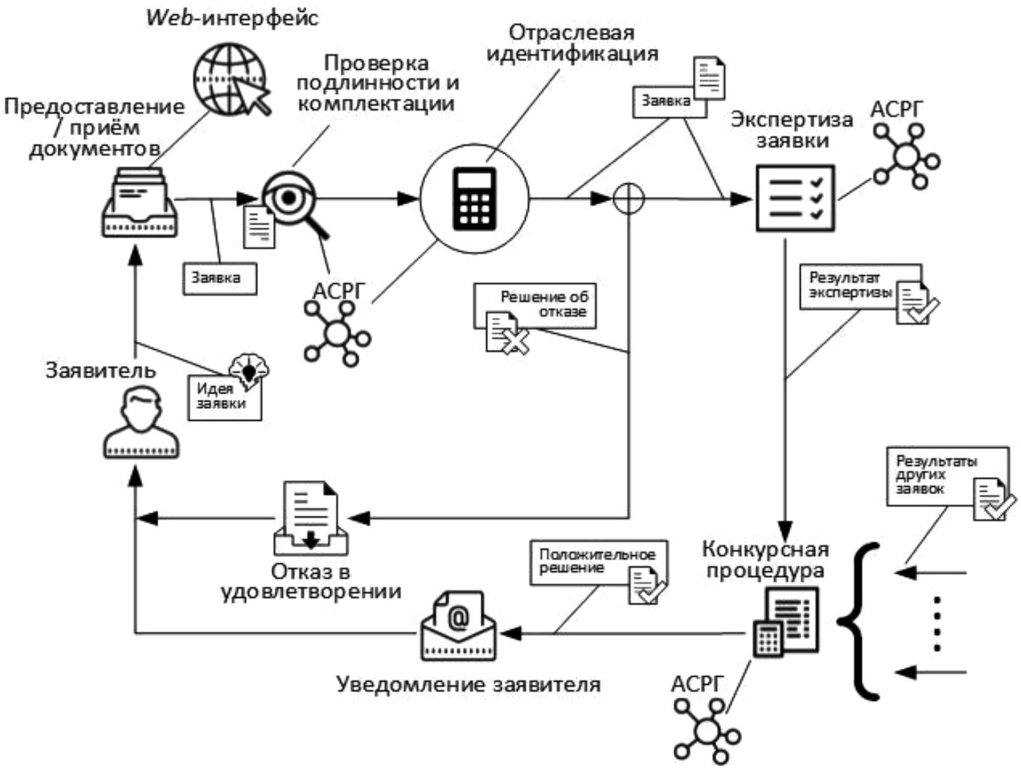

Рис. 3. Диаграмма поточной модели TO-BE работы АСРГ

Диаграмма поточной модели *TO-BE*, иллюстрирующая модель (5) для АСРГ, представлена на рис. 3.

Статическая модель оценки отдельной заявки может быть представлена выражением

$$\mathrm{H} \Rightarrow \left(\Psi, B_{\dot{E}_M}, B_{\ddot{E}_M}, f\right),$$

где f – функция свёртки обобщённого показателя, принимаемого к розыгрышу гранта. Соответственно модель розыгрыша может быть представлена следующим кортежем:

$$W = \left\langle \Psi, \mathrm{H}, G, w \right\rangle,$$

где G – множество грантов, w – функция выбора (конкурсная процедура).

Реализация модели (3) возможна только после определения отраслевой принадлежности заявки, для чего требуется соотнести поданную заявку с той или иной областью знания. Решение этой задачи возможно различными способами, в том числе методами тематического анализа, достаточно хорошо разработанными для различных документальных массивов.

Отраслевая идентификация может быть решена путём анализа соответствия содержания заявки неким признакам, например, отнесённых по областям знаний классификаторами ГРНТИ, УДК, ББК и др. Однако данный перечень не исчерпывает всех возможных признаков, поэтому реальная реализация системы может быть дополнена собственной, заранее разработанной, базой признаков, построенной на основе развешанных лингвистических ключей.

Предлагаемое решение не предусматривает машинного обучения и построения базы термов на основе анализа документов, как это применяется в общепринятой методологии *TF-IDF* (например, [2]) или в автоматизированных системах машинного обучения (например, [3] или [4]). Более того, мы предполагаем, что терминологическая база строго формализована для исключения неадекватного трактования предметной области заявки, как, например, предложено в [5].

Предположим, что сформирована база данных, содержащая некоторое множество L отраслей знаний, $|L| = n$, причём каждой области знания $l_i \in L$ поставлен в соответствие набор K_i признаков или ключевых слов $k_{ij} \in K_i, |K_i| = n_i$. Пусть R – множество поступивших заявок и $r_q \in R$. Относительно рассматриваемого множества R сопоставляем каждому признаку k_{ij} относительно заявки $r_q \in R$ величину w_{ij}^q, называемую

«весом». Кроме того, определяем частоты $f_{ij}^q = freq\left(k_{ij}, r_q\right)$ вхождений k_{ij} в r_q.

Далее выявляем значение частного показателя ind_i^q, устанавливающего степень соответствия области знания l_i заявке $r_q \in R$. Для этой цели возможно использование векторной модели (*vector space model*) метода *TF-IDF* [6] с учётом длины текста $\left|r_q\right|$:

$$ind_i^q = \frac{\sum_{j=1}^{n_i} f_{ij}^q w_{ij}^q}{\left|r_q\right|}. \quad (6)$$

Для компенсации прямого влияния длины текста на величину *ind* можно использовать модель, учитывающую усреднённую длину текста W_{avg} [5]:

$$\overline{ind_j^q} = \frac{ind_i^q}{ind_i^q + 0{,}5 + 1{,}5\frac{\left|r_q\right|}{W}}, \quad W = W_{avg}. \quad (7)$$

Модель (7) также не лишена недостатков, в частности не учитывает содержательного разнообразия, за счёт чего многократное повторение одного и того же термина может значительно повысить показатель *ind*, не имея при этом весомых оснований для отнесения заявки к определённой отрасли знания. Для компенсации этого недостатка может быть предложена следующая модель, основанная на уравнении Шеннона:

$$\overline{\overline{ind_i^q}} = ind_i^q \times \sum_{j=1}^{n_i}\left(-p_{ij}^q \log_2 p_{ij}^q\right), \quad p_{ij}^q = \frac{f_{ij}^q}{\left|r_q\right|}.$$

При определении отраслевой принадлежности возможен случай, когда показатель *ind* имеет одинаковое максимальное значение более чем для одной области. В этом случае возникает как бы условие конкуренции отраслей знаний за заявку r_q. Пусть для некоторых показателей p_α и p_β запись $H(p_\alpha; p_\beta)$ означает, что показатель p_β не превосходит показатель p_α. Условие конкуренции для отраслей l_α и l_β можно выразить следующим правилом:

$$\forall i \quad H\left(ind_\alpha^q; ind_i^q\right) \wedge H\left(ind_\beta^q; ind_i^q\right).$$

Выявляем множество

$$L_1 = \left\{ l_\alpha \in L \mid \forall i\, H\left(ind_\alpha^q; ind_i^q\right)\right\}.$$

Если $\left|L_1\right| = 1$, то заявка $r_q \in R$ должна быть сопоставлена области знания l_α. В случае, когда $\left|L_1\right| > 1$, формируем множество

$$L_2 = \left\{ l_\alpha \in L_1 \mid \forall i\, H\left(\overline{ind_\alpha^q}; \overline{ind_i^q}\right)\right\}. \quad (8)$$

Множество L_2 с одной стороны, с большей степенью вероятности позволяет выявить требуемую область знания l_α, с другой стороны, как и в модели (8), позволяет уточнить полученный результат, например

$$L_3 = \left\{ l_\alpha \in L_2 \mid \forall i\, H\left(\overline{\overline{ind_\alpha^q}}; \overline{\overline{ind_i^q}}\right)\right\}.$$

Выводы

Представленная модель отраслевой идентификации документа может быть использована для проведения предварительного этапа рассмотрения заявки при разработке универсальной автоматизированной системы распределения финансовых грантов. Данный подход также может быть применён к анализу текстов в любой области, при условии использования формализованных множеств термов, например, формализованной тематической рубрикации, классификации документов и пр. Результаты отраслевой идентификации будут использованы для формирования аналитического пакета инструментов, используемого в качестве ресурса для дальнейшего экспертного анализа заявки средствами искусственного интеллекта.

Список литературы

1. Яшин С.Н. Совершенствование инструментария системы оценки эффективности инновационных проектов, претендующих на получение грантов // Инновационное развитие. 2014. № 28 (604). С. 11–20.

2. 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. Т. 39. № 3. С. 429–438.

3. Царьков С.В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов // Естественные и технические науки. 2012. № 6. С. 456–464.

4. Попков М.И. Автоматическая система классификации текстов для базы знаний предприятия // International Journal of Open Information Technologies. 2014. Vol. 2. no. 7. С. 11–18.

5. Сироткин А.В., Шарыпов С.А. Тематическая модель рейтингования интернет-сайтов по критерию социальной значимости // Инженерный вестник Дона. 2016. № 4. [Электронный ресурс]. URL: http://ivdon.ru/ru/magazine/archive/n4y2016/3794 (дата обращения: 17.06.2019).

6. Бондарчук Д.В. Векторная модель представления знаний на основе семантической близости термов // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2017. № 3. [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/vektornaya-model-predstavleniya-znaniy-na-osnove-semanticheskoy-blizosti-termov (дата обращения: 17.06.2019).