Поиск научно-технических компетенций с использованием методик интеллектуального анализа текстов для формирования сообщества провайдеров решений в сфере открытых инноваций

О. П. Лукша,
старший консультант, председатель правления
o.luksha@rttn.ru

А. А. Наталенко,
к. ф.-м. н., руководитель отдела информационных технологий
a.natalenko@itp.rttn.ru

Г. Б. Пильнов,
к. т. н., исполнительный директор
g.pilnov@rttn.ru

А. Э. Яновский,
к. ф.-м. н., директор по проектам
a.yanovsky@rttn.ru

НП «Российская сеть трансфера технологий»

В последнее десятилетие появилась новая индустрия поставщиков услуг, которые помогают компаниям в реализации научно-технических проектов с использованием методологии открытых инноваций. Отдельно можно выделить так называемые акселераторы отрытых инноваций (АОИ), которые помогают клиентам получить прибыль путем подключения внешних партнеров (или провайдеров решений) ко всем стадиям инновационного проекта. Деятельность таких специализированных структур основана на совместном применении современных цифровых технологий и методологии краудсорсинга. При этом размер и структура сообщества провайдеров решений АОИ, а также соответствие компетенций его членов задачам проектов, рассматриваются в качестве ключевых параметров, определяющих его эффективность.

Авторами предлагается подход к формированию сообщества провайдеров решений, основанный на автоматизации задач поиска научно-технических компетенций персоналий на базе анализа исходной научно-технической задачи с использованием методов семантического анализа данных. Описывается архитектура и функционал программного комплекса, предназначенного для поиска информации и автоматизации бизнес-процессов АОИ.

Ключевые слова: открытые инновации, краудсорсинг, семантический анализ данных, провайдер решений, поиск технологической информации.

1. Модель краудсорсингового проекта и роль сообщества его участников

Мировой опыт демонстрирует широкое применение технологий краудсорсинга для решения задач в области маркетинга, управления, общественной экспертизы, социальной инженерии и некоторых других [1]. В последние 10 лет методы краудсорсинга стали также применяться в инновационных проектах для решения возникающих в них научно-технологических задач [2, 3].

Можно выделить 6 ключевых составляющих краудсорсингового проекта [4]:

1) формулировка проблемы, требующей поиска решения;
2) целевая аудитория — сообщество участников краудсорсингового проекта;
3) планируемый метод использования найденного решения;
4) методика и критерии оценки предложений;
5) инструменты для коммуникации и управления целевой аудиторией проекта (ИТ-платформа, мероприятия, маркетинг и т. д.);
6) награда за лучшее решение.

Целевая аудитория краудсорсингового проекта является одним из критически важных элементов, который во многом определяет его результативность [5, 6]. В зависимости от сферы реализации проекта могут оказаться существенными такие характеристики целевой аудитории как количество участников, их квалификация, регион проживания/работы, способность к генерации идей, к конструктивному обсуждению предложений и др. Для формирования релевантного со-

общества провайдеров решений и управления им необходимо:

- определить требования к квалификации и опыту участников — «портрет компетенций» потенциального участника сообщества;
- оценить минимальный критический размер сообщества;
- осуществить поиск потенциальных участников сообщества в соответствии с «портером компетенций» и требуемым размером сообщества;
- вовлечь и обеспечить мотивацию потенциальных членов сообщества к активному участию в краудсорсинговом проекте.

2. Сообщество провайдеров решений акселератора открытых инноваций

В акселераторах отрытых инноваций (АОИ) краудсорсинг используется для поиска новых решений научных, технологических или инженерных задач [2]. Поэтому целевую аудиторию здесь часто называют сообществом провайдеров решений (СПР) (solution providers). Как правило, участниками такого сообщества выступают исследователи, изобретатели, инженеры, а также команды стартап-проектов. Например, согласно исследованию своего сообщества компанией NineSigma по данным за 2014 г. [7] 62,4% участников имели ученую степень (PhD), 41,4% всех решений было предложено индивидуальными исследователями, а 58,6% — группами или коллективами.

Для эффективной работы АОИ к такому сообществу может предъявляться ряд специфических требований.

1. Значительный размер сообщества (от 20 тыс. до 1 млн человек). Это обусловлено тем, что как показывает статистика мировых бенчмарков, свои решения предлагают в среднем только 0,05-0,1% членов сообщества провайдеров решений, обладающих требуемой научно-технической квалификацией. В то же время клиент (постановщик задачи) АОИ обычно хочет видеть не одно предложение по решению, а несколько десятков (опционально даже сотен).
2. Разный тип участников: от индивидуальных изобретателей и исследователей до научных лабораторий и команд стартап-проектов. Такое требование связано с тем, что клиенты АОИ в качестве результата проекта могут ожидать получение как набора новых идей по преодолению технологической проблемы (idea generation), так и прототипа предлагаемого устройства или технологии.
3. Компетентность в научной или технической сфере. Предполагается возможность подтверждения квалификации участников сообщества имеющимися практическими достижениями. Это требование важно для отсечения «мусора» — изначально фантастических предложений, реализуемость которых объективно противоречит имеющимся знаниям в области науки и техники.
4. Готовность к продолжению сотрудничества по практической реализации предложенного решения (внедрению). Заказчики проектов в АОИ часто предусматривают возможность дальнейшей совместной работы (заказчика и провайдера лучшего решения) над доработкой и внедрением в деятельность заказчика найденного решения.
5. Географический охват. Участие провайдеров решений из разных регионов или стран может существенно расширить качество решений, обеспечить доступ к новым знаниям. Это может быть актуальным также, если заказчику необходимо не только решение, но и развитие сотрудничества в определенном географическом регионе и он заинтересован в партнерах.

При выполнении краудсорсингового проекта важна способность АОИ в короткие сроки (как правило, 1-2 мес.) распространить информацию о задаче и привлечь сообщество к ее решению. Поэтому важным конкурентным преимуществом АОИ является наличие инструментов, позволяющих быстро организовывать такую работу. Современные цифровые технологии (платформы, системы поиска) могут значительно облегчить решение этой задачи.

3. Возможности автоматизации задач по формированию сообщества провайдеров решений АОИ

Для управления сообществом в АОИ используются ИТ-платформы [3], которые автоматизируют бизнес-процессы АОИ и облегчают решение, как минимум, следующих задач:

- собственно формирование сообщества провайдеров решений;
- ведение реестра провайдеров решений и агрегация информации об их компетенциях;
- поддержание определенного уровня лояльности участников СПР и репутации АОИ в сообществе.

Таблица 1

Способы формирования сообщества провайдеров решений

Способ привлечения участников в сообщество	Особенности
Адресное приглашение к решению задачи или участию в сообществе провайдеров решений	Наиболее эффективный способ. Требует большой базы данных потенциальных провайдеров решений с характеристиками их компетенций и специализированной цифровой платформы для управления сообществом
Приглашение участников к решению задачи, распространяемое с использованием методов широкого маркетинга – рассылки, сайты, реклама, семинары и т. д.	Средняя эффективность. Хорошо работает в формализованных сообществах (исследователи в научных организациях, университетах и т. д.). Высокие затраты на проведение маркетинговых мероприятий
Персональные рекомендации/приглашения к участию в сообществе провайдеров решений	Хорошо работает в неформальных сообществах. Низкая скорость роста членов сообщества

Формирование сообщества провайдеров решений представляет собой нетривиальную задачу, которая решается различными способами (см. табл. 1).

Можно выделить два ключевых аспекта автоматизации процесса формирования СПР:

- создаваемые ИТ-решения должны позволять существенно сокращать время на формирование сообщества, одновременно обеспечивая «качество» его участников;
- для поиска информации о потенциальных участниках сообщества необходимо составление портрета компетенций участника сообщества. Автоматизация этой задачи требует выработки соответствующей методологии и алгоритмов.

Рассмотрим подробнее эти проблемы.

3.1. Ускорение формирования сообщества провайдеров решений

В отличие от проектов в социальной или общественной сфере, где свое решение может предложить практически любой участник сообщества, проекты в научно-технической сфере являются гораздо более сложными и требующими соответствующих компетенций, поэтому «плотность решений» значительно меньше. По этой причине размер СПР у АОИ рассматривается в качестве одного из параметров, определяющего эффективность АОИ и его конкурентоспособность. В среднем АОИ имеют сообщества порядка 20000 членов. АОИ, специализирующиеся на генерации идей или технических решений, часто имеют сообщества из 100000 участников и более. У ведущих мировых АОИ (Innocentive, NineSigma) размер сообществ достигает нескольких миллионов человек.

Анализ развития ведущих АОИ (Innocentive, NineSigma, Innoget и др.) показывает, что традиционный подход, связанный с естественным ростом СПР, требует значительного времени и существенных маркетинговых усилий. Рост сообщества до 1-3 млн чел. занимает 5-7 лет, сопровождается значительными финансовыми вложениями, что является существенным барьером для запуска и эффективного функционирования АОИ на начальном этапе.

Поэтому ИТ-платформа АОИ должна содержать инструментарий (специализированную поисковую систему), который может ускорить процесс формирования сообщества в сотни раз и сократить время на формирование релевантного технологическим задачам СПР до 1-2 недель, а также существенно снизить затраты на маркетинг АОИ.

3.2. Составление портрета компетенций участника сообщества

Провайдеры решений характеризуются своими компетенциями: наличием знаний и навыков (квалификации) и опытом решения аналогичных задач. Важны также определенные психологические характеристики личности. В целом портрет провайдера решения может быть описан 4 компонентами (примерно одинаковой важности):

- научно-техническая квалификация (базовые знания в предметной области);
- опыт решения аналогичных задач;
- способность предлагать новые идеи (нестандартные подходы);
- способность понимать и учитывать контекст, в котором решается задача (общая эрудиция).

Научно-техническая квалификация, в свою очередь, представляется как совокупность следующих характеристик:

- образование, в том числе наличие ученой степени;
- наличие печатных публикаций, участие в конференциях/выставках и т. п. мероприятиях;
- наличие изобретений (патентов);
- опыт участия в НИР и НИОКР проектах.

Исходя из перечисленных параметров, провайдеры решений, как правило, принадлежат к следующим группам:

- исследователи (например, ученые, работающие в научных центрах или университетах);
- инженеры (например, инженеры высокотехнологичных компаний, конструкторских бюро и т. д.);
- инновационные предприниматели, участники проектов по созданию стартап-компаний;
- индивидуальные изобретатели;
- инновационные посредники.

В общем случае в сообщество приглашаются провайдеры решений, обладающие компетенциями в различных областях науки и техники. Более сложный случай, но в то же время наиболее распространенный на практике, представляет собой специализированный поиск и привлечение в сообщество провайдеров для решения определенной конкретной задачи. В этом случае возникают конкретные требования к научно-технической квалификации потенциальных провайдеров, которые определяются из исходной задачи.

В данной ситуации сотрудник АОИ в первую очередь должен проанализировать исходный запрос и выделить оттуда перечень ключевых слов и фраз, которыми будет характеризоваться научно-техническая квалификация потенциальных провайдеров решений, требующихся для решения данной задачи. Таким образом будет сформирована основа портрета компетенций потенциального провайдера решений. Далее основа портрета компетенций должна быть дополнена информацией, исходя из структуры профиля компетенций. Дополнительной сложностью является то, что описание исходной задачи не всегда содержит достаточно информации для формирования релевантного портрета компетенций, что приводит к необходимости анализировать значительные объемы информации, возникающие в результате поиска по нечетко сформулированным критериям.

Составление портрета компетенций потенциального провайдера решений — это задача, которая требует не только высокой квалификации сотрудника АОИ, но и значительного времени для анализа и поиска похожих задач, из которых может быть выделена дополнительная полезная информация. Авторами, в рамках разработки ИТ-платформы для акселератора открытых инноваций, предлагается автоматизировать

процесс формирования (выделения из исходной задачи) портрета компетенций с использованием технологий семантического анализа данных, в частности, гибридного структурно-статистического алгоритма выделения ключевых слов и фраз на основе семантической сети. Подробно предлагаемый алгоритм описан далее.

3.3. Предлагаемый подход к автоматизации задач поиска научно-технических компетенций и формирования сообщества провайдеров решений

Предлагаемый подход, состоит в следующем:

1. На основе описания имеющейся задачи, требующей решения, с использованием семантического анализа формируется «портрет компетенций» потенциальных провайдеров решений.
2. С использованием сформированного «портрета компетенций» осуществляется семантический поиск в источниках информации, в которых публикуются данные, характеризующие компетенции физических лиц, принадлежащих к одной из групп потенциальных провайдеров решений.
3. Собранная информация используется для выявления конкретных персоналий и приглашения их к участию в решении задач.

Под источниками информации здесь понимаются базы данных, содержащие информацию, которая может характеризовать компетенции физических лиц/групп лиц, принадлежащих к одной из групп потенциальных провайдеров решений. Такой информацией являются:

- описания технологических запросов, предложений и иных информационных объектов, используемые в сетях трансфера технологий и технологических маркетплейсах;
- описания проектов, реализованных в рамках научно-технических конкурсов/грантов/программ поддержки и пр.;
- описания персоналий и их компетенций, используемые в профессиональных (тематических) сетях и сообществах;
- описания изобретений в патентных базах;
- научно-технические публикации.

Примерами источников данных, которые могут быть актуальны для поиска компетенций и формирования сообщества провайдеров решений, могут являться различные открытые ресурсы:

- профессиональные сети и сообщества (например, www.researchgate.net, www.linkedin.com и др.);
- сети трансфера технологий и технологические маркетплейсы (например, www.innocentive.com, www.innoget.com, www.autm.net);
- патентные базы (например, www.epo.org/searching-for-patents.html).

Исходное описание задачи, требующее поиска решения с привлечением сообщества провайдеров, равно как и описания различных информационных объектов в приведенных источниках, являются слабоструктурированными текстовыми описаниями и не содержат в явном виде перечня требуемых компетенций, а модели данных, используемые в различных информационных источниках, значительно отличаются друг от друга. Возникает задача сравнения содержания документов и поиска семантически подобных информационных объектов.

4. Использование латентно-семантический анализа для разработки алгоритма поиска информации о компетенциях провайдеров решений

Для программной реализации поиска семантически подобных информационных объектов могут использоваться различные методы семантического анализа. Начиная от самых простых статистических метрик на базе TF-IDF [8] (Okapi BM25 и подобные), методов снижения размерности векторного пространства терминов (LSI, LPI) [9], их вероятностных вариаций (DFR, pLSA, LDA) [10], до сложных методов хэширования и нейросетевых методов (LSH, staked-RBM, boostin-SSC и др.) [11]. Одним из достаточно простых в реализации, давно известных и хорошо себя зарекомендовавших методов является латентно-семантический анализ (LSA) [12]. LSA предназначен для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных. В основе LSA лежат принципы факторного анализа, и этот метод широко используется для классификации или кластеризации больших объемов текстовых документов, учитывая контекстно-зависимые значения слов.

Программная реализация метода LSA осуществляется по следующему алгоритму:

1) индексация исходного текста (формирование перечня слов и фраз (терминов) из исходного текста, используемых для дальнейшей работы алгоритма);
2) формирование вероятностной матрицы индексированных терминов;
3) сингулярное разложение матрицы;
4) снижение размерности вероятностной матрицы;
5) анализ полученных результатов.

Ключевым этапом в методе LSA является этап индексации текста, когда из текста выделяется набор ключевых слов и фраз, которые формируют «индекс терминов», описывающих семантическое ядро текста.

На сегодняшний день известно множество методик и алгоритмов выделения ключевых слов и фраз (статические, структурные (графовые), нейросетевые, гибридные) [13]. Но до настоящего времени не разработана единая общепринятая методика выделения ключевых слов. В связи с этим работы в данном направлении продолжаются и однозначного лучшего алгоритма на сегодняшний день не существует. Наряду с обязательным использованием тривиальных процедур удаления стоп-слов и стемминга [14], совершенствованием статистических и структурных методов (TextRank, Rake, DegExt и др.) в последние годы с ростом вычислительных мощностей акцент разработчиков сместился в сторону более ресурсоемких гибридных и нейросетевых решений, в том числе обучаемых на основе текстовых корпусов алгоритмов. Большинство алгоритмов выделения ключевых слов

и фраз обладают множество скрытых настраиваемых параметров. Такие параметры подбираются эмпирическим путем на основе обучения алгоритма на некотором базовом эталонном корпусе текстов, который наилучшим образом описывает лингвистические особенности текстов.

При индексировании информационных объектов в рамках задачи по поиску и формированию сообщества провайдеров решений возникают дополнительные сложности, которые затруднительно преодолеть с помощью известных методов:

1. Описание одной технологической задачи не всегда содержит достаточное количество слов и фраз, которые могут в должной мере охарактеризовать потенциального провайдера решений. Как показывает опыт, анализа одного проекта недостаточно для того, чтобы сформировать полноценный портрет будущего провайдера решений. Требуется более широкая аналитика по пулу подобных проектов и уже найденных решений (в зависимости от специфичности задачи это могут быть десятки и сотни подобных проектов).
2. При индексировании описания технологической задачи необходимо учитывать возможные междисциплинарные связи. Первоначально сформированный индекс должен дополняться, учитывая возможности (вероятность) решения поставленной задачи за пределами предметной области, описание которой сформулировано в самой задаче.
3. При анализе научно-технических текстов (как самой технологической задачи, так и информационных объектов из релевантных источников) более важны не отдельные слова, а словосочетания или фразы. В текстах может использоваться различная терминология (слова синонимы, аббревиатуры, термины на других языках).

Для преодоления этих сложностей мы предлагаем использовать структурно-статистический алгоритм выделения ключевых слов и фраз на базе семантической сети.

5. Гибридный структурно-статистический алгоритм выделения ключевых слов и фраз с использованием семантической сети

Дополнительная информация для преобразования начального индекса терминов может быть получена из семантической сети. Семантическая сеть, требующаяся для поддержки процесса индексирования информационных объектов, должна содержать термины и информацию о вероятностях связей между терминами внутри определенной предметной области и о междисциплинарных связях. Используя данную семантическую сеть, можно выполнить более точное и гибкое индексирование исходного описания технологической задачи, подготавливая данные для дальнейшей реализации поиска подобных информационных объектов методом LSA. Первоначально полученный из исходного текста индекс терминов, будет преобразовываться: изменяться (используя слова синонимы, аббревиатуры, термины на других языках) и дополняться (путем образования словосочетаний и фраз) для достижения результата поиска с требуемой точностью и в зависимости от «семантического качества» исходного текста и текстов в источниках поиска. Используя семантическую сеть, исходный индекс терминов может дополняться терминами в рамках предметной области задачи и терминами в рамках известных междисциплинарных связей.

Размерность индекса терминов может варьироваться как в большую, так и в меньшую сторону с целью получения наилучших результатов поиска. Приведем иллюстрацию предлагаемого алгоритма (рис. 1).

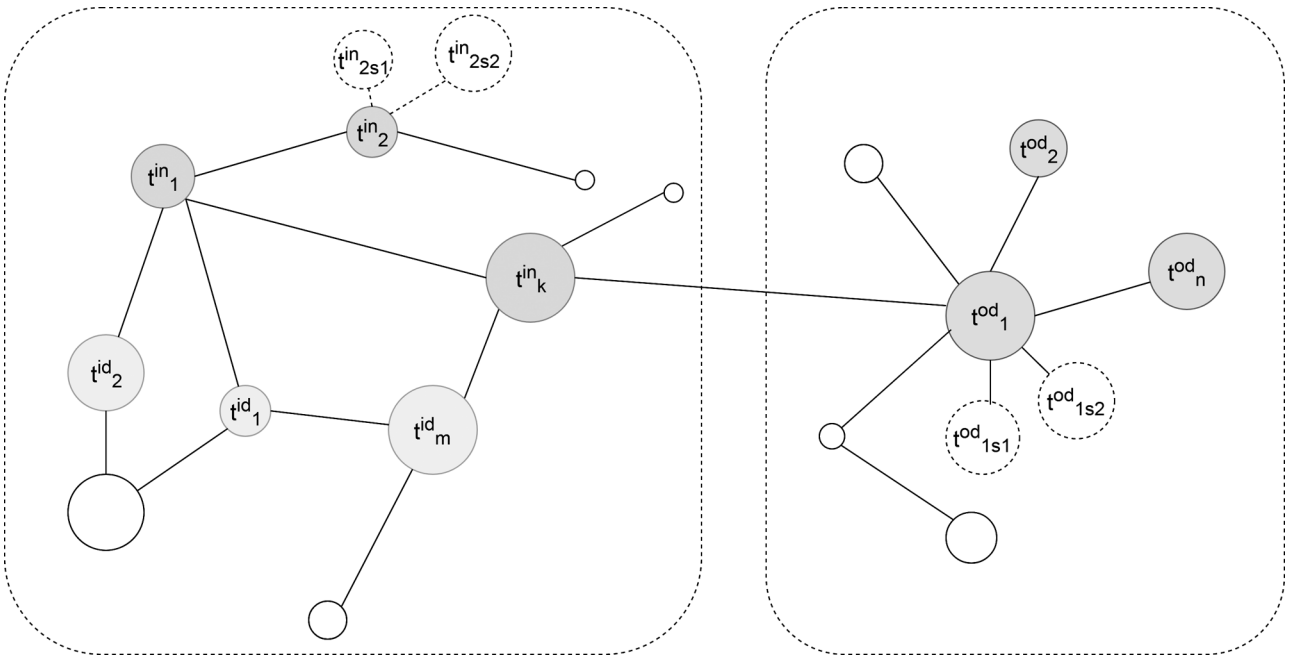

Рис. 1. Гибридный структурно-статистический алгоритм выделения ключевых слов и фраз с использованием семантической сети (общая схема)

Введем обозначения:

- $T_{\text{in}}\langle t_1^{\text{in}}...t_k^{\text{in}}\rangle$ — множество терминов, полученных из исходного текста;
- $T_{\text{id}}\langle t_1^{\text{id}}...t_m^{\text{id}}\rangle$ — множество терминов, входящих в предметную область задачи, но не содержащихся в исходном тексте;
- $T_{\text{od}}\langle t_1^{\text{od}}...t_n^{\text{od}}\rangle$ — множество терминов, не входящих в предметную область задачи и не содержащихся в исходном тексте.

Преобразование терминов осуществляется по следующему алгоритму:

1. Каждый элемент t_k^{in} множества $T_{\text{in}}\langle t_1^{\text{in}}...t_k^{\text{in}}\rangle$ проверяется на предмет вхождения в семантическую сеть. Если элемент отсутствует, то он в дальнейшем может быть рассмотрен в качестве кандидата на включение в семантическую сеть (этот вопрос рассматривается отдельно).
2. Если элемент присутствует в сети, то для него проверяется наличие синонимов

$$T_{\text{in}\,s}\langle t_1^{\text{in}} - t_{1s1}^{\text{in}} ... t_k^{\text{in}} - t_{ks1}^{\text{in}}\rangle.$$

   Если синонимы существуют, то они могут быть добавлены к исходному элементу или исходный элемент может быть заменен на синонимы в зависимости от вероятностей их использования.
3. Для каждого элемента t_k^{in} проверяется наличие аббревиатур, терминов на других языках, вероятность вхождения в устойчивые словосочетания и фразы.
4. В случае существования соответствующих элементов они добавляются к исходному множеству.
5. Для каждого элемента t_k^{in} проверяется наличие связей с другими элементами в рамках предметной области и вероятности этих связей.
6. Добавление элементов множества $T_{\text{id}}\langle t_1^{\text{id}}...t_m^{\text{id}}\rangle$ к элементам множества $T_{\text{in}}\langle t_1^{\text{in}}...t_k^{\text{in}}\rangle$ может осуществляться в зависимости от вероятности конкретной связи, либо по расстоянию по графу (например, добавляются все ближайшие элементы, с которыми существуют связи).
7. Для каждого элемента t_k^{in} и t_m^{id} проверяется наличие связей с другими элементами за границами предметной области.
8. Добавление элементов множества $T_{\text{od}}\langle t_1^{\text{od}}...t_n^{\text{od}}\rangle$ к элементам множества $T_{\text{in}}\langle t_1^{\text{in}}...t_k^{\text{in}}\rangle$ осуществляется в зависимости от вероятности конкретной связи.

В результате выполнения алгоритма исходное множество терминов $T_{\text{in}}\langle t_1^{\text{in}}...t_k^{\text{in}}\rangle$ преобразуется в множество

$$T_{f\text{in}}\langle t_1^{\text{in}}...t_k^{\text{in}}\rangle\langle t_1^{\text{id}}...t_m^{\text{id}}\rangle\langle t_{1s2}^{\text{od}}...t_n^{\text{od}}\rangle,$$

которое далее используется в LSA.

Таким образом, дополнительное использование семантической сети терминов позволяет решить проблемы, связанные с особенностями индексирования текстов в задаче формирования сообщества провайдеров решений, которые не могут быть решены традиционными методами обработки текстов, применяемыми в методе LSA.

Для построения базовой (начальной) семантической сети необходимо использовать корпус научно-технических документов различных тематик (максимально охватывающий предметные области деятельности АОИ). Обновление и расширение семантической сети может осуществляться в процессе работы АОИ за счет данных из используемых при поиске информационных источников.

7. Прототип программного комплекса для поиска научно-технических компетенций персоналий и формирования сообщества провайдеров решений

Для практической реализации предлагаемого подхода к формированию сообщества провайдеров решений должен быть создан программный комплекс (ПК). Для работы с семантической сетью могут использоваться современные Фреймворки (например, Apache Jena [15] или Eclipse RDF4J [16]). Что касается программной реализации самого метода LSA, включая тривиальные методы предварительной обработки текстов (удаление стоп-слов, стемминг и пр.), то он может быть реализован на базе таких программных платформ и библиотек, как Apache Lucene [17], Sphinx [18], Xapian [19] и др., которые реализуют базовый функционал поисковой системы и могут быть расширены и адаптированы для решения конкретной задачи.

В рамках верхнеуровневого представления архитектура ПК может быть описана совокупностью 5 основных функциональных подсистем (рис. 2):

1) подсистема получения текстов из источников и первичной обработки (включая модуль взаимодействия с API источников (SOAP/REST), модуль парсинга веб-сайтов (DOM/CSS, средства быстрой настройки), модуль загрузки пользовательских запросов (задач));
2) подсистема индексирования текстов (модуль подготовки текста и начального индексирования и модуль, реализующий гибридный структурно-статистический алгоритм преобразования начального индекса на основе данных семантической сети (онтологии));
3) подсистема LSA (сравнение индексированных текстов методом латентно-семантического анализа);
4) подсистема агрегации и представления результата;
5) подсистема формирования семантической сети (тематической модели) из извлекаемых текстов.

Функциональные подсистемы ПК будут использоваться в двух процессах, которые могут выполняться как независимо друг от друга, так и одновременно:

Первый процесс связан с задачей загрузки и индексирования текстов, формированием тематической модели и наполнением семантической сети. В этом процессе используются подсистемы получения текстов из источников и первичной обработки (рис. 2, блок 1), индексирования текстов (рис. 2, блок 2) и формирования онтологии (рис. 2, блок 5).

Процесс представлен на рис. 3.

Алгоритм формирования онтологии:

- используя подсистему индексирования текстов, из каждого текста выделяется индекс терминов (используется тривиальная процедура без дополнительного обогащения за счет тематической модели);

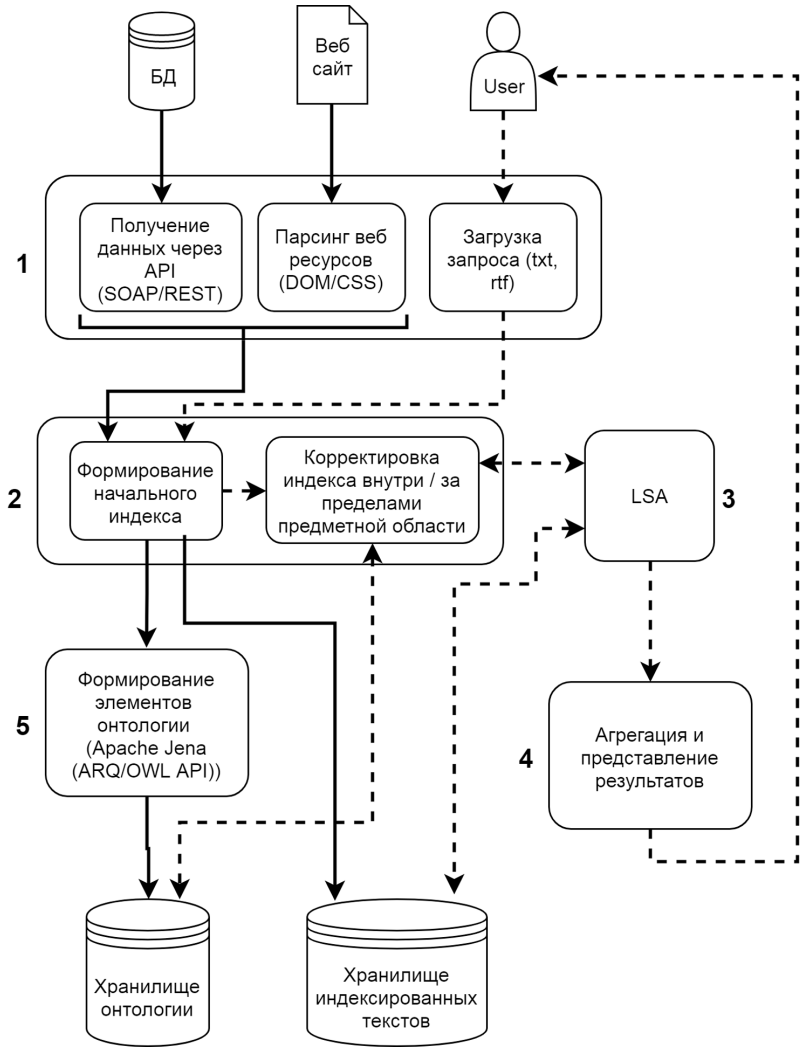

Рис. 2. Верхнеуровневое представление архитектуры программного комплекса для формирования сообщества провайдеров решений

- для каждого термина создается онтологический класс (OWL/RDF модель) [20], для каждого класса создаются объектные свойства (InverseObjectProperty) со всеми классами в рамках обрабатываемого индекса, каждое свойство помечается весовым коэффициентом равным 1;
- в случае если класс уже присутствует в онтологии, то для него просто добавляются объектные свойства;
- в случае если объектное свойство уже присутствует в онтологии, то для него весовой коэффициент увеличивается на 1.

В такой модели онтология не только хранит информацию о связях между терминами, но и информацию о вероятностях связей того или иного термина, что позволяет строить масштабируемые тематические модели, путем вариаций вероятностей. Онтология хранится в RDF модели в триплстore.

Второй процесс связан с задачей непосредственно поиска провайдеров решений, в этом процессе используются подсистемы индексирования текстов (рис. 2, блок 2), LSA (рис. 2, блок 3) и агрегации и представления результата (рис. 2, блок 4). Процесс представлен на рис. 4.

Рассмотрим его более подробно. Отбор релевантных информационных объектов — текстов, характеризующих компетенции провайдеров решений, осуществляется по следующему алгоритму:

1) загрузка текста задачи. Анализ неструктурированного описания проекта. Выделение перечня ключевых слов (формирование начального индекса терминов);

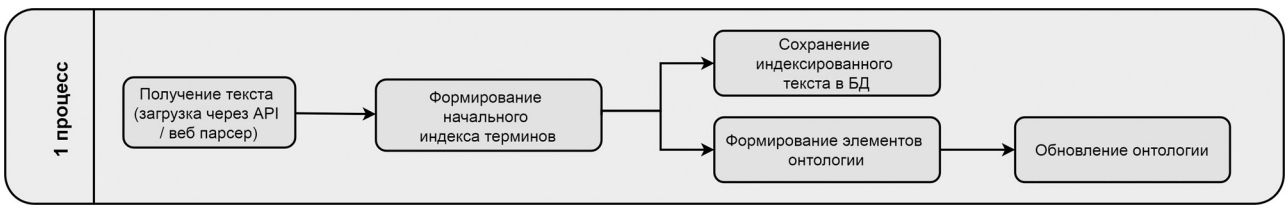

Рис. 3. Алгоритм формирования тематической модели и наполнения семантической сети

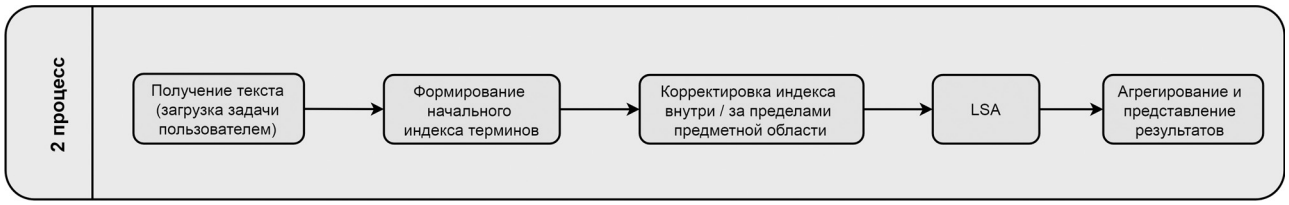

Рис. 4. Алгоритм поиска провайдеров решений

2) уточнение и дополнение перечня ключевых слов и построение ключевых фраз, используя семантическую сеть (связи внутри предметной области);
3) уточнение и дополнение перечня ключевых слов и фраз, используя семантическую сеть (известные связи между терминами, которые относятся к разным предметным областям);
4) формирование и распространение поискового запроса по актуальным источникам информации (поиск методом LSA);
5) обработка (агрегация) результата и представление в форме, удобной для дальнейшего анализа.

Используя ПК, пользователь сможет через единый пользовательский интерфейс получать информацию о наличии семантически подобных (соответствующих исходному запросу) информационных объектов (профилей провайдеров решений) из десятков релевантных информационных источников (тематических сетей/маркетплейсов/краудсорсинговых площадок).

Описанный алгоритм будет реализован и апробирован в рамках прототипа программного комплекса акселератора открытых инноваций, создаваемого НП «Российская сеть трансфера технологий».

Заключение

Автоматизация процесса формирования сообщества провайдеров решений позволит существенно сократить время, необходимое на поиск и систематизацию информации о потенциальных провайдерах решений.

Использование предлагаемого программного комплекса создает дополнительные конкурентные преимущества для запуска и работы АОИ, а именно:

1. Возможность быстро сформировать сообщество провайдеров решений. У известных АОИ на это ушли годы работы. Традиционные методы маркетинга (встречи, семинары, публикации) здесь оказываются очень затратными и недостаточно эффективными.
2. Возможность формировать сообщество провайдеров решений для конкретной технологической задачи на основе заданных слабо формализованных критериев и описания контекста, в рамках которого решается задача.
3. Снижение затрат на содержание большого числа штатных или привлекаемых экспертов для решения задач поиска и анализа научно-технической информации.

Список использованных источников

1. Daren C. Brabham. Crowdsourcing. MIT, 2013
2. F. Piller, K. Diener. Brokers and Intermediaries for Open Innovation — A Global Market Study. 2013.
3. О. П. Лукша, А. А. Наталенко, Г. Б. Пильнов, А. Э. Яновский, Акселераторы открытых инноваций на основе информационных платформ//Инновации. № 12. 2017.
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5362025.
5. https://irevolutions.org/2010/05/05/towards-a-model-for-successful-crowdsourcing.
6. О. Манчулянцев. Открытые конкурсы как источник инновационных идей//Открытые инновации для крупных компаний. Сб. статей. М.: Московская школа управления Сколково, 2011.
7. http://www.ninesigma.com/File%20Library/Infographics/SP-Survey_infographic.pdf.
8. A. Rajaraman et al. Mining of Massive Datasets. Cambridge University Press, 2011. P. 1-17.
9. I. Vinnarasi Tharania et al. Improved Correlation Preserved Indexing For Text Mining//IJIRCCE. Vol. 2. Issue 1. 2014. P. 2482-2490.
10. T. Hofmann Thomas. Probabilistic latent semantic indexing//In Proc. of the SIGIR 1999. P. 50-57.
11. D. Zhang. Extensions to Self-Taught Hashing: Kernelisation and Supervision, The SIGIR 2010 Workshop on Feature Generation and Selection for Information Retrieval (FGSIR), 2010.
12. T. K. Landauer. An Introduction to Latent Semantic Analysis// Discourse Processes. Vol. 25. 1998. P. 259-284.
13. K. S. Hasan. Automatic keyphrase extraction: a survey of the state of the art//In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. P. 1262-1273.
14. M. F. Porte, An algorithm for suffix stripping//Program. Vol. 14. № 3. 1980. P. 130-137.
15. http://jena.apache.org.
16. http://rdf4j.org.
17. http://lucene.apache.org.
18. http://sphinxsearch.com.
19. https://xapian.org.
20. OWL 2 Web Ontology Language, W3C Recommendation, 2012. https://www.w3.org/TR/owl2-overview.

Search for scientific and technical competencies using the methods of intellectual text analysis to form a community of solution providers in the field of open innovation

O. P. Luksha, senior consultant, board chairman.

A. A. Natalenko, candidate of physico-mathematical sciences, chief information officer.

G. B. Pilnov, PhD, managing director.

A. E. Yanovsky, PhD, project director.

(Russian technology transfer network)

In the last decade, a new industry of service providers has emerged that helps companies to realize scientific and technical projects using the methodology of open innovation. Special group of them are so-called «accelerators of open innovation» (AOI), which bring benefits for clients by connecting external partners (or solution providers) to all stages of an innovative project. The activities of such specialized structures are based on the joint application of modern digital technologies and the methodology of crowdsourcing.

The size of the AOI' solution providers community is considered as one of the critical factor that determines the effectiveness of the AOI and its competitiveness. The authors suggest an approach for the formation of a solution providers community based on automating the search for scientific and technical competences of personalities on the basis of the analysis of the initial scientific and technical problem using methods of semantic data analysis. The paper describes the architecture and functions of software designed to search information on solution providers and AOI' business process automation.

Keywords: open innovations, semantic data analysis, crowdsourcing, solution providers, technology scouting.