УДК 004.622

Сбор и подготовка данных для целей создания IT профиля регионов

Костенников Д.В. (ИТМО)

Научный руководитель – доктор экономических наук, профессор Максимова Т.Г. (ИТМО)

Введение. Исследование IT профиля региона является актуальным и важным в контексте современной экономики и развития технологий. В России постепенно происходит переход от классической промышленности к цифровой экономике. Цифровая трансформация является национальной целью, и от качества и темпов ее реализации в регионах зависит развитие страны [1]. И для успешной реализации этого процесса важно иметь полное и информированное представление о IT потенциале каждого региона.

В наши дни, благодаря быстрому прогрессу технологий, анализ данных является важным инструментом для определения основных тенденций и оценки цифрового потенциала региона. Патентная активность, а также разработка программ для электронных вычислительных машин и баз данных являются одними из основных областей, определяющих IT профиль.

Основная часть. В опубликованных открытых данных Роспатента были найдены два реестра. Первый — «Открытый реестр программ для электронно-вычислительных машин» [2], а второй — «Открытый реестр баз данных» [3]. Общий объем данных составляет 195188 строк, где каждая строка содержит информацию об одной регистрации программы для ЭВМ или базы данных, включая название правообладателя, название программы или базы данных и регистрационный номер.

В результате анализа названия правообладателей, было выявлено, что в них всегда присутствуют слова, являющиеся общими признаками для различных типов организаций, такие как "ОАО", "ФГБОУ", "Российская Федерация от имени которой выступает". Количество таких слов ограничено. Для классификации правообладателей были созданы 10 списков ключевых слов, с помощью которых в названиях выделено указание на тип организаций, такие как компания, государственная организация, учебное заведение и научное учреждение, частное учебное заведение, среднее специальное учебное заведение, некоммерческая организация и физическое лицо. Для большинства правообладателей, в названии которых есть слова из этих списков, эти слова были удалены, оставляя только название организации. Это предотвращает дублирование организации в датасете при его агрегации. Затем данные были агрегированы по правообладателям, чтобы датасеты имели вид: название правообладателя (строка), количество зарегистрированных объектов (целое число), регистрационные номера (список), название программы для ЭВМ/БД (список), тип организации (строка). Таким образом удалось уменьшить общий объем данных до 44301 строк без потери информации.

Для задачи определения региона регистрации компаний был использован API сервиса "DaData", который позволяет по названию организации получить адрес ее регистрации [4]. С помощью этого сервиса были найдены города регистрации для 15792 компаний из 19248, регионы – для 16091 и федеральные округа – для 16079. Эти данные были добавлены в датасеты, дополнив их колонками город (строка), регион (строка), федеральный округ (строка).

Правообладатели, относящиеся к госсектору и имеющие в названии регион, например "Санкт-Петербург от имени которого выступает комитет по информатизации и связи", были классифицированы с учетом их региональной принадлежности. Для этого был создан словарь со всеми регионами Российской Федерации и различными вариантами их написания. Таким образом, для 483 правообладателей из 955 удалось определить регион.

Выводы. Были собраны и обработаны данные по регистрации программ для ЭВМ и БД. Удалось уменьшить изначальный объем с 195188 до 44301 строк. Из них для 17384 был найден регион. В дальнейшем необходимо найти регионы для правообладателей с типом учебное заведение и научное учреждение, частное учебное заведение, среднее специальное учебное заведение, некоммерческая организация, и на основании этих данных создать интерактивную карту объектов интеллектуальной собственности Российской Федерации.

Список использованных источников:

1. Абрамов В. И., Андреев В. Д. Цифровая экосистема региона: практические аспекты реализации и структурные компоненты //Ars Administrandi (Искусство управления). – 2023. – С. 251-271.
2. Открытый реестр программ для электронно-вычислительных машин [Электронный ресурс], URL: https://rospatent.gov.ru/opendata/7730176088-evm
3. Открытый реестр баз данных [Электронный ресурс], URL: https://rospatent.gov.ru/opendata/7730176088-bd
4. DaData [Электронный ресурс], URL: https://dadata.ru/api/suggest/party/