МОНИТОРИНГ КОЛЛЕКТИВНЫХ СРЕДСТВ РАЗМЕЩЕНИЯ НА ОСНОВЕ ДАННЫХ ИЗ ОТКРЫТЫХ ИСТОЧНИКОВ

Ю. В. Пестова, О. А. Николайчук, Д. Е. Косогоров, А. И. Павлов
Институт динамики систем и теории управлений имени В. М. Матросова СО РАН
Иркутск

В работе представлены онтологическое представление информации о коллективных средствах размещения и их услугах, методы сбора, обработки и визуализации данных, разработаны интерактивные информационные панели, обеспечивающие анализ и интерпретацию собранных данных. В качестве источников данных были выбраны «Островок.ру», «101 Отель» и «Мир Турбаз», из которых осуществлялся сбор данных посредством метода парсинга. Собранные данные обработаны, преобразованы и интегрированы в единый вид, затем выполнена идентификация объектов размещения. Выполнен анализ данных посредством разработанных панелей визуализации с помощью возможностей BI-платформ, результаты анализа интерпретированы: определены территории с наибольшим количеством средств размещения, их рейтинг, множество наиболее популярных услуг и др.

Ключевые слова: туризм, мониторинг, открытые данные, онтология, сбор данных, визуализация, средства размещения

MONITORING OF COLLECTIVE ACCOMMODATION FACILITIES BASED ON DATA FROM OPEN SOURCES

Yu. V. Pestova, O. A. Nikolaichuk, D. E. Kosogorov, A. I. Pavlov
Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences
Irkutsk

The paper presents an ontological representation of information about collective accommodation facilities and their services, methods for collecting, processing and visualizing data, and interactive dashboards have been developed that provide analysis and interpretation of the collected data. As data sources, "Ostrovok.ru", "101 Hotel" and "Mir Turbaz" were chosen, from which data was collected using the parsing method. The collected data is processed, transformed and integrated into a single form, then the identification of accommodation facilities is performed. The data analysis was carried out using the developed visualization panels using the capabilities of BI-platforms, the results of the analysis were interpreted: the territories with the largest number of accommodation facilities, their rating, many of the most popular services, etc. were determined.

Keywords: tourism, monitoring, open data, ontology, data collection, visualization, accommodation facilities

В Стратегии развития туризма в Российской Федерации на период до 2035 года отмечено, что для роста конкурентоспособности и раскрытия потенциала туристского продукта необходимо обеспечить повышение доступности актуальных отраслевых данных со стороны участников туристского рынка и развитие цифровых платформ продвижения туристских продуктов и брендов, цифровых средств навигации и формирования туристского продукта, а также разработать систему мониторинга качества оказываемых услуг на приоритетных туристских территориях [1].

Официальные статистические данные о состоянии региональной сферы туризма Байкальского региона не отражают полную информацию о средствах размещения, объектах питания, оказываемых услугах и их качестве и т.д. [2, 3, 4]. Среди формируемых статистических показателей отметим: численность размещенных лиц, численность ночевок, доходы коллективных средств размещения и др. [5]. Один из главных недостатков этих данных – отсутствие их географической привязки для обеспечения возможности оценки рекреационной нагрузки рассматриваемой территории.

Целью данной работы является разработка методики сбора данных о коллективных средствах размещения и оказываемых ими услугах, осуществление сбора и визуализации данных Байкальской природной территории для обеспечения дальнейшего мониторинга территориальной сферы туризма.

Автоматизация разрабатываемой методики обеспечит:

– оперативное получение информации о коллективных средствах размещения обеспечит формирование туристского профиля территории, мониторинг и районирование территории;

– информационную поддержку для принятия решений малым и средним предпринимательством в сфере туризма: определение территории для бизнеса и выбор, обоснование и формирование туристических продуктов для бизнеса;
– информационную поддержку для принятия решений клиентов туристических услуг: подбор территории для отдыха и туристических продуктов.

В качестве источников открытых данных выбраны: сайты-агрегаторы объектов размещения: «Островок.ру», «101 Отель» и «Мир Турбаз». На основе запроса «Байкал» и полученного в результате списка страниц формируется база мест пребывания, номерного фонда, туристических услуг, их стоимости, а также оценок пользователей на основе рейтинга и комментариев.

Анализ источников информации позволил сформировать онтологию коллективных средств размещения (рис. 1) для дальнейшей унификации понятий, используемых в разрабатываемых информационных моделях.

Рис. 1. Онтология коллективных средств размещения

Услуги, предоставляемые в коллективных средствах размещений, также имеют детализированную структуру, которую можно описать через онтологическое представление (рис. 2).

Список источников для сбора данных может быть расширен, но их структура однозначно сформирована на основе представленной онтологии.

Для получения такой информации применен метод сбора данных: парсинг для извлечения данных из динамических источников из-за отсутствия API. Программными средствами языка программирования Python, библиотеки Selenium, были написаны алгоритмы, которые осуществляют такие действия с помощью агента в окне браузера по каждому веб-ресурсу. Библиотека Selenium отличается от других программных средств сбора данных: с ее помощью, можно совершать динамическое изменение разметки сайта без отправления запросов к ресурсу. При применении парсинга следует учитывать следующие аспекты:

– разработанный алгоритм является уникальным для каждого веб-ресурса, он полностью зависит от его разметки и необходимости совершаемых действий пользователем для извлечения необходимой информации,
– в рамках одного источника данных путь к информации (на основе разметки сайта), записанный в программном коде, может быть изменен с течением времени,
– алгоритм должен содержать задержки выполнения кода для имитации действий настоящего пользователя и во избежание нагрузки на сайт, что может спровоцировать его нестабильную работу,
– для повышения скорости сбора данных используется распараллеливание потоков, где для каждого создается агент.

Статус парсинга не обозначен на законодательном уровне, однако из анализа нормативно-правовых актов можно выделить категории информации, которые нельзя нелегально собирать таким методом – это информация, обладающая статусом банковской, налоговой, государственной, коммерческой или являющаяся профессиональной тайной. Также запрещено использовать парсинг в целях получения материальной и иной выгоды, искусственного создания неправомерной конкуренции и получения закрытых персональных и иных данных, охраняемых законодательством Российской Федерации [6]. Необходимо отметить, что используемые в работе источники данных не соответствуют перечисленным аспектам нормативно-правовой базы РФ.

Собранные данные подвергаются предобработке данных, после чего требуется однозначно идентифицировать средства размещений по всем источникам. Сложность заключается в том, что записи по одному объекту имеют существенно различные названия и значения других свойств (ошибка в координатах, разная стоимость, категория мест и т.п.). И, наоборот, разные объекты могут иметь одно название. Однозначно идентифицируются те объекты, что имеют в записях одинаковые названия без учета типа объекта (гостиница,

база отдыха и др.) и находятся в одном населенном пункте. Для остальных осуществляется поиск ближайшей записи по нормированным показателям дистанций: Левенштейна по названию, названию населенного пункта, адреса и косинусное расстояние между координатами.

Рис. 2. Онтология категорий туристических услуг

Идентификация позволяет агрегировать количественные и качественные показатели для получения более достоверной информации, на основе нескольких источников. Например: усреднение стоимости средств размещения, пользовательской оценки (рейтинга), суммирование количества отзывов, категорий и популярных услуг. Из 1163 записей идентифицированы 685 объектов.

Результаты анализа отражены посредством визуализации данных с помощью инструментальных средств BI-платформы. BI-платформы – инструменты бизнес-аналитики, с помощью которых возможно объединение данных из различных источников, их обработка и анализ посредством создания отчетов – дашбордов, которые представляют собой интерактивные информационные панели с помощью технологий визуализации. Панели изменяются при действии фильтров, установки параметров и позволяют интерпретировать результаты анализа. В рамках работы сформированная модель данных обеспечивает:

- отображение средней минимальной цены и количества объектов по населенным пунктам/районам в виде комбинированной диаграммы;
- отображение услуг и их детализации по количеству мест размещений, которые их предлагают, для выявления востребованных в виде столбчатой диаграммы и цветной таблицы;
- отображение коллективных средств размещения на карте в виде плотности точек по количеству номеров;
- отображение ключевых общих показателей количества мест и услуг;
- возможность фильтрации по населенным пунктам/районам, рейтингу мест и услугам.

Наибольшее количество объектов размещения (152) сосредоточено в поселке Хужир (остров Ольхон). На втором месте – поселок Листвянка (101 объект). Самый высокий средний чек наблюдается у средств размещений в поселках Мангутае и Новоснежной, расположенных в южной части Байкала, в этих поселках также самое низкое число объектов размещения, данные факты можно интерпретировать наличием низкой конкуренции и отсутствием привлекательных туристических услуг, что можно использовать при формировании решений по развитию новой туристской рекреационной зоны. Анализ услуг 339 объектов размещения с рейтингом «Великолепно» (рис. 3) показал, что наиболее распространены услуги категории «развлечения и спорт», а среди подкатегорий – «пешие экскурсии». Такие данные могут быть использованы при создании нового бизнеса и планировании его перечня услуг.

В результате работы сформировано онтологическое представление о коллективных средствах размещения и их услугах, что позволяет однозначно определить структуру данных при сборе из открытых источников. В качестве источников данных были выбраны «Островок.ру», «101 Отель» и «Мир Турбаз», из которых осуществлялся сбор посредством метода парсинга.

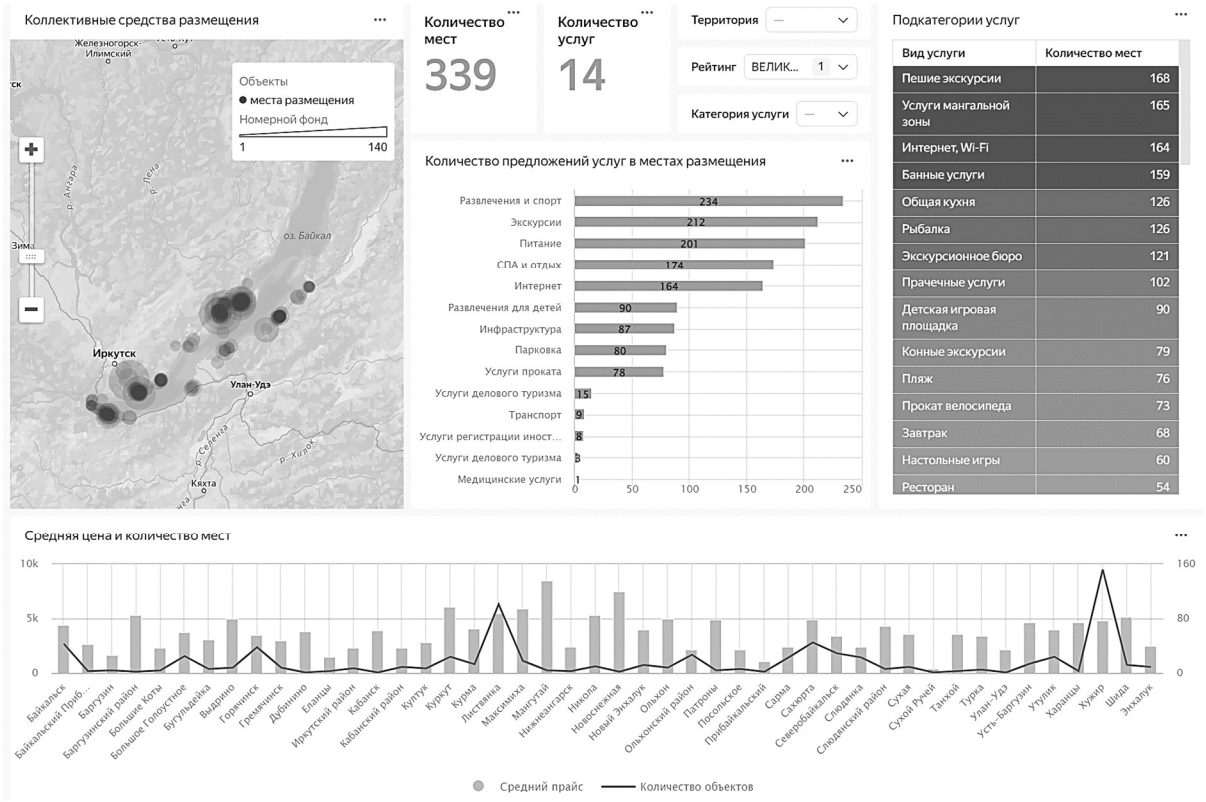

Рис. 3. Фрагмент информационной панели с фильтрацией по рейтингу – «Великолепно»

Собранные данные обработаны и преобразованы в единый вид, после чего выполнена идентификация мест для анализа объектов. Результаты анализа интерпретированы посредством разработанных панелей визуализации с помощью возможностей BI-платформ. В дальнейшем планируется более детальное изучение полученных данных.

Исследование выполнено за счет гранта Российского научного фонда № 23-28-00844 «Мониторинг сферы регионального туризма на основе анализа данных из открытых источников» (https://rscf.ru/project/23-28-00844/).

ЛИТЕРАТУРА

1. Постановление Правительства Российской Федерации «О Стратегии развития туризма в России до 2035 года». г. Москва, от 20 сентября 2019 года № 2129-р [Электронный ресурс]. URL: static.government.ru/media/files/FjJ74rYOaVA4yzPAshEulYxmWSpB4lrM.pdf (дата обращения: 22.03.2023).
2. Котельников Д. А. Формирование системы показателей для ведения мониторинга устойчивого развития туристских территорий // Конкурс научных инноваций: перспективы развития науки в современном мире. Сборник статей по материалам всероссийского научно-исследовательского конкурса. Уфа, 2020. С. 41-50.
3. Лебедева Ю. А. Организация мониторинга качества туристских услуг на муниципальном уровне. Чебоксары, 2020.
4. Шмидт Ю. Д., Рубцова Н. В. Формирование системы мониторинга эффективности функционирования сферы туристско-рекреационных услуг региона // Материалы V Всероссийской научно-практической конференции «Интеллектуальный и ресурсный потенциалы регионов: активизация и повышение эффективности использования» / Под науч. ред. А. П. Суходолова, Н. Н. Даниленко, О. Н. Баевой. Иркутск: Байкальский государственный университет, 2019. С. 520-524.
5. Туризм. Федеральная служба государственной статистики [Электронный ресурс]. URL: https://rosstat.gov.ru/statistics/turizm (дата обращения: 12.04.23).
6. Дятлова Е. В., Янгличева Ю. Р. Парсинг и закон // Вестник экономики, права и социологии. 2022.№ 2. С. 49-52.