

О ПОДХОДАХ К РЕАЛИЗАЦИИ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ ИСТОРИЧЕСКИХ ПАМЯТНИКОВ РЕСПУБЛИКИ БЕЛАРУСЬ

Л.В. Рудикова-Фронхёфер, А.И. Жвалецкий

*Гродненский государственный университет имени Янки Купалы,
ул. Ожешко, 22, 230023, г. Гродно, Беларусь
lada.rudikowa@gmail.com, zhvaleuski.andrei@gmail.com*

В статье приведены основные теоретические и практические аспекты разработки информационно-аналитической системы исторических памятников и достопримечательностей Республики Беларусь с использованием технологии контейнеризации, принципов предметно-ориентированного проектирования и шаблона разделения ответственности команд и запросов. В разработке используется нейронная сеть BERT от Google для обработки запросов пользователей на естественном языке и последующего формирования релевантных результатов. Описывается процесс построения системы, а также приводятся основные характеристики разработки.

Ключевые слова: веб-приложение; анализ данных; исторические памятники; семантический поиск.

ON APPROACHES TO THE IMPLEMENTATION OF THE INFORMATION AND ANALYTICAL SYSTEM OF HISTORICAL MONUMENTS OF THE REPUBLIC OF BELARUS

L.V. Rudikova-Fronhoefer, A.I. Zhvaleuski

*Yanka Kupala State University of Grodno,
Ozheshko str., 22, 230023, Grodno, Belarus,
lada.rudikowa@gmail.com, zhvaleuski.andrei@gmail.com*

The article presents the main theoretical and practical aspects of the system necessary for the development of an information and analytical system of historical monuments in the Republic of Belarus using containerization technologies, the principles of subject-oriented design and a template for the division of responsibility of teams and requests. The development uses Google's BERT neural network to process user requests in natural language and then produce relevant results. The process of building the system is described, the characteristics of the application are given.

Keywords: web application; data analysis; historical monuments; semantic search.

Введение

Проектирование и разработка универсальной информационно-аналитической системы предусматривает сбор, обработку и анализ данных об исторических памятниках, достопримечательностях и других ценных объектах Республики Беларусь. Обработка собранных данных, их анализ и структурированное представление позволит упростить процесс изучения и временные затраты на поиск информации об исторических памятниках и достопримечательностях. Аналитическая система предполагает также открытый интерфейс в виде веб-API для взаимодействия с данными, позволяя производить интеграцию с внешними системами.

Цель исследования – выбор эффективных технологий и инструментов для создания информационно-аналитической системы исторических памятников и достопримечательностей в Республике Беларусь. Предлагаемая разработка будет предоставлять средства для взаимодействия с информацией, расширенного поиска и анализа хранимых данных.

Задачи исследования: провести анализ имеющихся инструментов разработки; выявить требования и функциональность, предъявляемые к системе; выбрать оптимальный набор средств разработки; спроектировать архитектуру системы; разработать модель данных; разработать программное решение для информационно-аналитической системы (серверная и клиентская часть); провести тестирование и отладку системы; сделать выводы и подготовить техническую документацию.

1. Методология исследования / теоретические основы

Объектами исследования являются данные об исторических памятниках, а также связанными с ними места, объекты, события, личности и т.д. Предмет исследования – алгоритмы сбора и обработки исторических данных, визуализация, анализ и их использование. Методами исследования являются структурная методология проектирования и объектно-ориентированный анализ.

Научная новизна и практическая значимость обусловлена разработкой программных решений, алгоритмов сбора, обработки и визуализации данных информационно-аналитической системы исторических памятников и достопримечательностей Республики Беларусь. В дальнейшем разработка может использоваться в образовательном процессе, а также в туристических целях, предоставляя структурированную информацию об исторических памятниках,

связанных событиях и исторических личностях, а также документацию и модели проектирования систем подобного рода

2. Результаты и их обсуждение

Рассмотрим некоторые аспекты предметной области, связанной с историческими памятниками и достопримечательностями. Историко-культурный объект характеризуется следующими параметрами: название; описание; фото- и видео- материалами; классификацией по критериям всемирного наследия ЮНЕСКО: культурные (I, II, III, IV, V, VI), природные (VII, VIII, IX, X); категория историко-культурной ценности Республики Беларусь (0, 1, 2, 3, А, Б и без категории); место расположения. В зависимости от типа историко-культурного объекта, ему могут быть назначены различные свойства. Например, материал изготовления, скульптор, высота скульптуры для памятника, или конфессия, архитектурный стиль для храма.

Историко-культурный объект является основной сущностью в системе при проведении инфологического моделирования предметной области исследования [1]. Однако в системе будет храниться и информация связанного плана (например, местоположение, география места, исторические события, связанные с местом и т.д.), а также информация о посещениях данного объекта и мероприятиях, связанных с данным объектом. Кроме того, следует также учитывать и взаимосвязи всех выделенных объектов и ограничения, накладываемые предметной областью. На рисунке представлен фрагмент концептуальной модели, связанной с объектами исторической ценности и их основными свойствами [2].

Что касается подходов к реализации, предлагаемая система разрабатывается на основе микросервисного подхода, а также с использованием технологий контейнеризации и оркестрации, что значительно упрощает процесс внедрения изменений в систему и её тестирование. В данном случае используются Docker и Kubernetes, а для управления инфраструктурой (запуск, обновление контейнеров и т.д.) – Terraform [3].

Таким образом, проект разрабатывается на основе различных технологий, которые позволяют решать поставленные задачи наилучшим образом. Часть подсистем, отвечающих за реализацию бизнес-логики, не связанной с анализом данных, реализована на платформе ASP.NET и языке программирования C#. Данные хранятся в базе данных PostgreSQL [4]. Для построения графического пользовательского интерфейса используется Next.js фреймворк для React и библиотека компонентов

Materiel UI. Подсистема, отвечающая за анализ данных, реализуется на языке Python с применением его различных библиотек, например, TensorFlow, OpenCV, pandas.

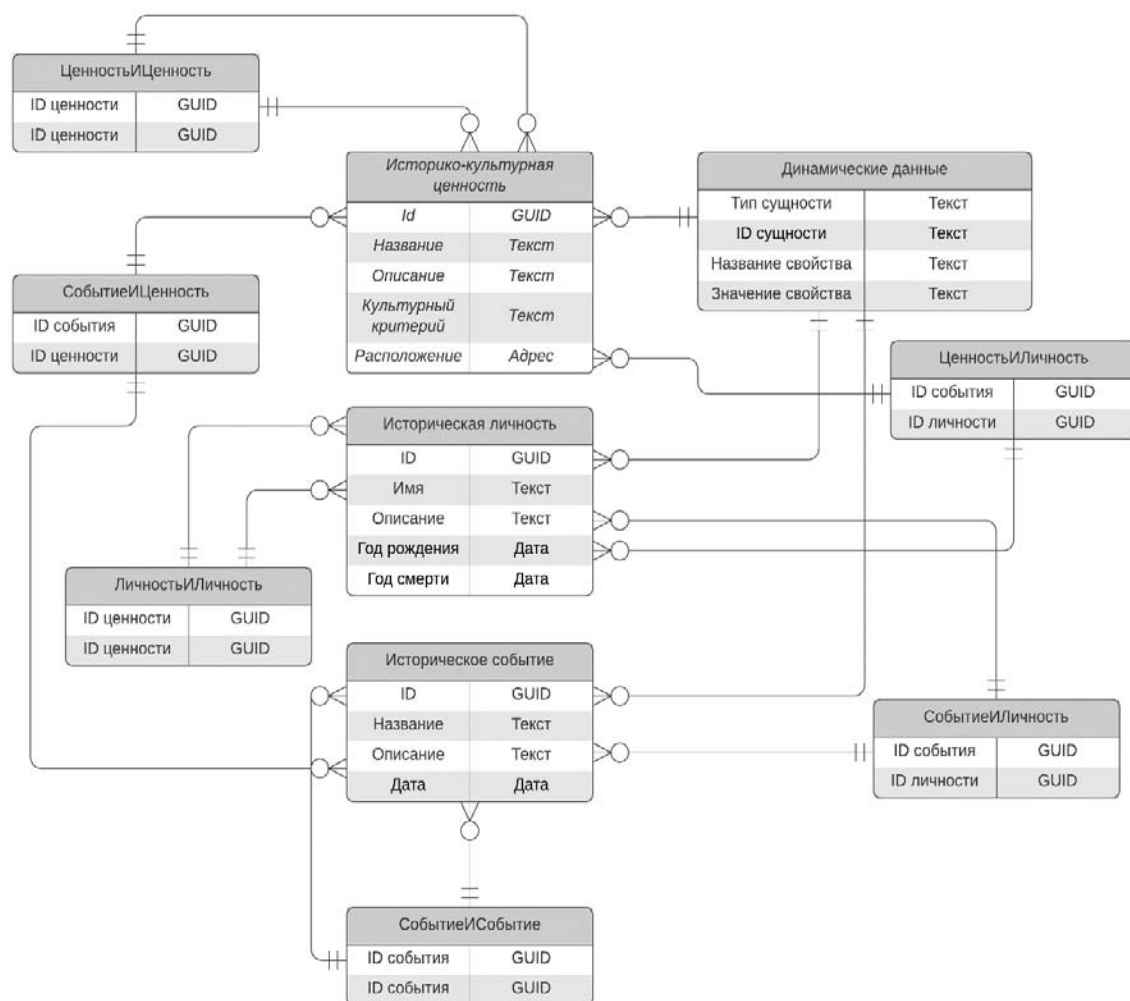


Рисунок – Фрагмент концептуальной модели, связанной с объектами исторической ценности

Основой подсистемы анализа исторических данных является алгоритм обработки запросов на естественном языке. Каждый запрос проходит через обученную модель и, в результате, выдает список данных (события, исторические личности, памятник), которые соответствуют запросу пользователя. Для решения задачи обработки естественного языка используется языковая модель Bidirectional Encoder Representations from Transformers (BERT) от Google [5].

BERT использует Transformer, attention механизм, который изучает контекстуальные отношения между словами (или вложенными словами) в тексте. В своей vanilla форме Transformer включает в себя два отдельных

механизма — кодер, который считывает вводимый текст, и декодер, который выдает прогноз для задачи. Поскольку целью BERT является создание языковой модели, необходим только механизм кодирования. Подробная работа Transformer описана в статье Google [6]. В отличие от направленных моделей, которые считывают вводимый текст последовательно (слева направо или справа налево), преобразовательный кодер считывает всю последовательность слов сразу. Эта характеристика позволяет модели изучать контекст слова на основе всего его окружения (слева и справа от слова). При обучении языковых моделей возникает проблема определения цели прогнозирования. Многие модели предсказывают следующее слово в последовательности (направленный подход), который, по сути, ограничивает контекстное обучение. Чтобы преодолеть эту проблему, BERT использует две стратегии обучения, которые приводятся ниже.

Masked LM (MLM). Перед подачей последовательностей слов в BERT 15% слов в каждой последовательности заменяются токеном [MASK]. Затем модель пытается предсказать исходное значение замаскированных слов на основе контекста, предоставляемого другими, не замаскированными словами в последовательности.

Next Sentence Prediction (NSP). В процессе обучения BERT модель получает пары предложений в качестве входных данных и учится предсказывать, является ли второе предложение в паре последующим предложением в исходном документе. Во время обучения 50% входных данных представляют собой пару, в которой второе предложение является последующим предложением в исходном документе, в то время как в других 50% в качестве второго предложения выбирается случайное предложение из корпуса. Предполагается, что случайное предложение будет отсоединено от первого предложения.

Заключение

Таким образом, предлагаемая разработка может представлять, как коммерческую ценность, так и академическую пользу. Авторам работы видится дальнейшая перспектива применения разработанного продукта в роли самостоятельного сервиса с возможностью его дальнейшего развития и внедрения в системы, связанные с обработкой информации различного профиля.

Библиографические ссылки

1. Рудикова Л.В. Проектирование баз данных: Учебное пособие для студентов высш. учеб. заведений по специальностям «Программное обеспечение информационных технологий», «Экономическая кибернетика», «Прикладная математика (научно-педагогическая деятельность)», «Информационные системы и технологии (в экономике)». Минск: ИВЦ Минфина, 2009. 352 с.
2. Рудикова Л.В. О системе хранения и обработки информации о произведениях художественной ценности на основе технологии складирования данных // Проблемы современной экономики: глобальный, национальный и региональный контекст: сб. науч. ст. / ГрГУ им. Я. Купалы; редкол.: М.Е. Карпицкая (гл. ред.), С.Е. Витун (зам. гл. ред.) [и др.]. Гродно: ГрГУ, 2021. С. 210–216
3. Эванс Э. Предметно-ориентированного проектирования (DDD): структуризация сложных программных систем: Пер. с англ. М.: ООО «И.Д. Вильямс». 2011. 448 с.
4. Backend for Frontend Security Framework Overview (Duende IdentityServer) [Электронный ресурс]. URL: <https://docs.duendesoftware.com/identityserver/v6/bff/overview/> (дата обращения: 20.08.2022.)
5. BERT Explained: State of the art language model for NLP [Электронный ресурс]. URL: <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270> (дата обращения: 28.08.2022.)
6. Attention Is All You Need [Электронный ресурс]. URL: <https://arxiv.org/pdf/1706.03762.pdf> (дата обращения: 11.08.2022.)