УДК 004.6-024.11:004.738.5

СИСТЕМА АНАЛИЗА ДАННЫХ ТЕМАТИЧЕСКИХ САЙТОВ

Батура М.П., Пилецкий И.И., Волорова Н.А.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь, bmpbel@bsuir.by

Аннотация. Рассмотрены принципы построения и архитектура системы комплексного анализа данных интернет-источников, приведены примеры использования системы в учебном процессе.

Ключевые слова. BIG DATA, RDF-схема, машинное обучение, графовые базы данных, графовые алгоритмы.

В современном обществе основным источником получения информации является интернет; по прогнозам специалистов к 2025 году цифровое взаимодействие через интернет вырастет до 181 зеттабайт [1]. Однако, колоссальный объем данных в сети и простота их получения порождают и проблему избыточности данных, поиска необходимой, наиболее полезной и достоверной информации. Задача анализа данных является актуальной и ее актуальность возрастает с развитием информационных технологий и включает в себя не только получение достоверной информации, но и сокращение времени для ее получения и анализа.

На кафедре информатики БГУИР на протяжении нескольких лет выполнялись работы по разработке программных комплексов анализа данных из интернет-источников и реализации Системы комплексного анализа данных тематических сайтов – ИСКАД ИИ.

Большинство известных ИТ компаний разрабатывают или имеют аналитические средства анализа данных и последние данные Gartner [2] трендов в области ИТ показывают возрастающую роль графовых технологий в области данных и аналитики.

Одним из способов представления знаний является применение специальных глобальных словарей предметных областей, мета-описаний и специальных языков. Многие важные мировые сайты, используют специальную технику описания ресурса Resource Description Framework (RDF, «среда описания ресурса») [3].

RDF представляет собой модель обмена данными, которая описывает, как данные сериализуются и как обмениваются ими. Использование такого подхода позволяет выполнять описание знаний в тематических предметных словарях и обмениваться этими знаниями с другими сайтами. Данные для импорта и экспорта на таких RDF сайтах могут быть представлен в нескольких форматах (JSON-LD, Turtle, N-Triples, RDF/XML, TriG и N-Quads, TriG *).

Такой подход описания сайтов позволяет применить специальную методологию построения (генерации) графовой БД из описания RDF данных. Тематическая графовая БД, содержит базу знаний сайта в виде графа знаний, что позволяет применять различные аналитические алгоритмы ML для более глубокого анализа данных сайта.

Предлагаемый подход был апробирован и протестирован при разработке многоцелевого модифицируемого кластера для анализа данных интернет источников с выявления наиболее важных публикаций в некоторой предметной области, тематического анализа этих публикаций, выявлении лидера научного направления и предсказания тенденций развития направлений и взаимодействия групп людей.

Для разработки системы было принято мультиплатформенное решение для анализа данных различных предметных областей, с помощью которого можно быстро получать данные и информацию о предметной области с различных мировых крупных систем, на базе быстрого построения графовой(ых) базы данных предметной области и выполнять более глубокий анализ.

Основными компонентами системы (ИСКАД ИИ) являются: компонент получения данных с интернет источников; компонент графовая БД и граф знаний; компонент извлечения свойств из графовой БД и их анализ с помощью алгоритмов ML; компонент интеграции, который представляет собой специальный веб-сайт ИСКАД ИИ.

Общая функциональная архитектура ИСКАД ИИ приведена на рисунке 1.

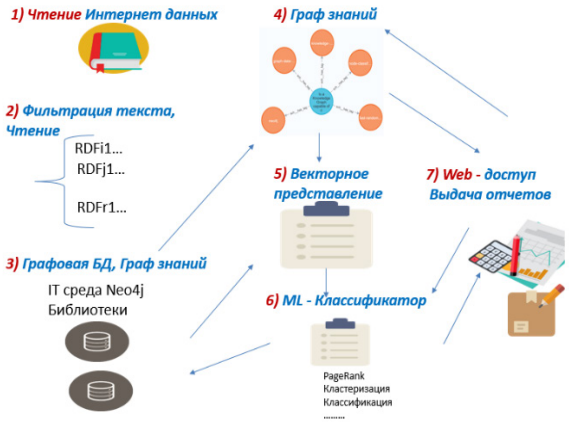

Рисунок 1 – Общая функциональная архитектура ИСКАД ИИ

Компонент получения данных с интернет источников выполняет различные операции над данными, включая очистку, структурирование, нормализацию и приведение к единому формату. Работа компонента выполняется в специальной ИТ среде, основными компонентами которой являются: графовая СУБД Neo4j Desktop, специальные библиотеки (плагины) расширяющие возможности анализа данных в графовой БД и фреймворки для разработки веб-сайта ИСКАД ИИ. Такое решение обеспечи-

вает быстрое построение тематических графовых БД использующих RDF описание данных в различных форматах.

Компонент графовая БД и граф знаний реализуются с помощью ИТ среды и графовой БД Neo4j (рисунок 1). Данная БД обладает свойствами горизонтального масштабирования, с ростом данных не деградирует, работает в значительно быстрее чем реляционная БД и обеспечивает требования ACID (*atomicity, consistency, isolation, durability*). База Данных Neo4j обеспечивает работу с миллионами узлов и отношений и доступ к данным как на языке запросов Cypher, так и на популярных языках программирования. Использование БД Neo4j дает возможность получать графы знаний и графически визуализировать наборы данных и результаты запросов. В процессе доступа скачивания данных с сайтов компонентом получения данных, графовая БД строится и модифицируется автоматически.

Компонент извлечения свойств из графовой БД, может использовать технологию включений (embedding), что позволяет построить векторы свойств меньшей размерности для более глубокого анализа данных. Компонент извлечения свойств из графовой БД и их анализ с помощью алгоритмов машинного обучения позволяет реализовывать запросы пользователей, выдавать тематические графы знаний и обеспечивать анализ данных в графовых БД преобразовывая свойства в узлах и ребрах (отношениях) в векторное представление с целью применения для дальнейшего анализа данных.

Компонентом интеграции является веб-сайт ИСКАД ИИ который обеспечивает работу компонентов системы и предоставляет доступ к получению аналитических данных пользователям системы. Веб-сайт позволяет пользователю создавать, обновлять и взаимодействовать с тематическими графовыми базами данных, а также выполнять запрос к графу знаний для получения и выдачи репортов

Зарегистрированный пользователь может иметь доступ к следующим сервисам:

– просмотру публикаций различных предметных областей; поиску наиболее цитируемых авторов предметной области;

– просмотру параметров некоторой предметной области;

– просмотру различной информации об авторе публикации.

Пользователь может видеть имя, количество публикаций и сами публикации автора, а также гистограмм по авторам и статьям. Гистограммы по статьям и авторам строятся с помощью алгоритма PageRank.

Администратор Веб-сайта может добавлять, удалять пользователей и предоставлять им расширенные права, управлять состоянием базы данных. Администратор может менять структуру БД, содержимое БД и делать замену одной БД на сайте на другую.

Веб-сайт является центральным компонентом, который облегчает управление данными, взаимодействие пользователей и получение информации из системы. Для реализации компонента была использована клиент-серверная архитектура. Такой подход обеспечивает интеграцию и эффективность в работе с пользователями и компонентами системы.

Данное решение и разработанная методология реализованы как система комплексного анализа данных, которая развернута в учебной лаборатории кафедры информатики и используется в качестве материалов лекционного курса и примеров построения систем и анализа данных по дисциплине «Модели и методы обработки больших объёмов данных», для второй ступени высшего образования а также используются в разделах программы NoSQL базы данных, графовые базы данных, в качестве примеров моделирования и построения физических и социальных сетей, обработки и анализа публикаций интернет источников.

Литература

1. DOMO [Электронный ресурс]. – Электронные данные. – Режим доступа: https://www.domo.com/data-never-sleeps

2. Gartner [Электронный ресурс]. – Электронные данные. – Режим доступа: https://www.gartner.com/en/newsroom/press-releases/2021-03-16-gartner-identifies-top-10-data-and-analytics-technologies-trends-for-2021/

3. Среда Описания Ресурса (RDF): Понятия и Абстрактный Синтаксис [Электронный ресурс]. – Электронные данные. – Режим доступа: https://www.w3.org/2007/03/rdf_concepts_ru/Overview.html/ The RDF Data Cube Vocabulary (w3.org)

SYSTEM FOR DATA ANALYSIS OF THEMATIC SITES

M.P. Batura, I.I. Piletsky, N.A. Volorova

Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus, bmpbel@bsuir.by

Abstract. The principles of construction and architecture of a system for complex analysis of data from Internet sources are considered, and examples of using the system in the educational process are given.

Keywords. BIG DATA, RDF schema, machine learning, graph databases, graph algorithms.