

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

(повне найменування вищого навчального закладу)

Факультет математики, фізики та інформаційних технологій

(повне найменування інституту, назва факультету (відділення))

Кафедра математичного забезпечення комп'ютерних систем

(повна назва кафедри (предметної, циклової комісії))

Дипломна робота

на здобуття ступеня вищої освіти «магістр»

(освітньо-кваліфікаційний рівень)

на тему:

Аналіз вподобань методами Data Mining на картографічних
даних

Analysis of preferences using Data Mining methods based on
cartographic data

Виконав: студент денної форми навчання

спеціальності 123 – Комп'ютерна інженерія

(шифр і назва напрямку підготовки, спеціальності)

Єремєєв Андрій Юрійович

(прізвище, ім'я, по-батькові)

Керівник к.ф.-м.н., доц., Т.І.Петрушина

(науковий ступінь, вчене звання, прізвище та ініціали, підпис)

Рецензент д.т.н., проф., Є.В. Малахов

(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент к.т.н., М.Г.Глава

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:

Протокол засідання кафедри

№ від « » 2020 р.

Завідувач кафедри

Є.В. Малахов

(підпис)

(прізвище, ініціали)

Захищено на засіданні ЕК №

протокол № від « » 2020 р.

Оцінка

(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

Н.Ф. Казакова

(підпис)

(прізвище, ініціали)

Одеса – 2020

АНОТАЦІЯ

Дипломна робота присвячена темі «Аналіз вподобань методами Data Mining на картографічних даних».

В рамках даної теми досліджуються методи аналізу вподобань людей при виборі місць проведення часу. У зв'язку з сучасними тенденціями проведення часу людей в закладах харчування, освіти та розваги, необхідність в пошуку місць відпочинку займає одну з центральних місць в повсякденному житті людини.

Метою роботи є підвищення якості рекомендацій місць відпочинку на основі аналізу вподобань відпочиваючих.

В результаті даної роботи запропонована гнучка архітектура, заснована на виконанні фонових процесів, яка містить функціонал для обробки даних, зберігання всієї необхідної інформації в тому числі про об'єкти відвідування, відгуків про них і користувачів. Запропоновані засоби збору, очистки та трансформації даних і виконана апробація запропонованих рішень у вигляді реалізації фонових процесів для виконання цього функціоналу. Передбачена можливість гнучкого додавання нового джерела даних в розробленому спеціальному набору механізмів виконання ресурсномістких обчислювальних процесів. Моніторинг стану процесів запропоновано здійснювати за допомогою спеціального графічного інтерфейсу. Також реалізований процес комунікації основної системи зі кластером Spark для зберігання даних і навчання моделі.

Результатом роботи є модель рекомендаційної системи, побудована методом матричної факторизації, фонові процеси для її підтримки і засоби надання рекомендацій за допомогою цієї моделі у вигляді REST API.

АННОТАЦИЯ

Дипломная работа посвящена теме «Анализ предпочтений методами Data Mining на картографических данных».

В рамках данной темы исследуются методы анализа предпочтений людей при выборе мест времяпровождения. В связи с современными тенденциями проведения времени людей в заведениях питания, образования и развлечения, необходимость в поиске мест отдыха занимает одну из центральных мест в повседневной жизни человека.

Целью работы является повышение качества рекомендаций мест отдыха на основе анализа предпочтении отдыхающих.

В результате данной работы предложена гибкая архитектура, основанная на выполнении фоновых процессов, которая содержит функционал для обработки данных, хранения всей необходимой информации в том числе про объекты посещения, отзывах них и пользователей. Предложены способы сбора, очистки и трансформации данных, и выполнена апробация предложенных решений в виде реализации фоновых процессов для выполнения этого функционала. Предусмотрена возможность гибкого добавления нового источника данных в разработанном специальном наборе механизмов выполнения ресурсоемких вычислительных процессов. Мониторинг состояния процессов предложено выполнять с помощью специального графического интерфейса. Также реализован процесс коммуникации основной системы с кластером Spark для хранения данных и обучения модели.

Результатом работы является модель рекомендательной системы, построенной методом матричной факторизации, фоновые процессы для ее поддержки и способы предоставления рекомендаций с помощью этой модели в виде REST API.

ABSTRACT

The subject of master's thesis is "Analysis of preferences using Data Mining methods based on cartographic data".

Within the framework of this subject, the methods of analyzing people's preferences during the selection of the places for leisure are investigated. In connection with the modern tendencies of how people spend their time in the places for eating, education and entertainment, the need to find places for leisure takes one of the central roles in the daily life.

The goal of the work is to improve the quality of recommendations for places for leisure based on the analysis of people preferences.

As a result of this work, the flexible architecture, which is based on the background processes, was introduced. It includes the functionality for data processing, storing all the necessary information, including the data about objects for visit, reviews about them and users. The ways to collect, clean and transform data were proposed and the evaluation of the suggested solution was conducted in the form of the implementation of background processes executing this functionality. The ability to add a new data source in a flexible way was provided using the developed special set of background heavyweight computational processes execution mechanisms. The possibility to monitor the state of the processes using a special graphical interface was added to the overall functionality as well as the processes for the communication between the main system and Spark cluster to store the data and model training.

The result of the work is a recommendation system model, built using the Matrix Factorization method, background processes for its maintenance and means to provide recommendations using this model with the help of REST API.

ЗМІСТ

СПИСОК СКОРОЧЕНЬ.....	7
ВСТУП.....	8
1 ОГЛЯД МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ РЕКОМЕНДАЦІЙНИХ СИСТЕМ	11
1.1 Основні концепції Data Mining.....	11
1.2 Основний підхід до вирішення завдань інтелектуального аналізу даних.....	12
1.3 Класифікація задач методів інтелектуального аналізу даних і рекомендаційних систем	13
1.3.1 Основні методи Data Mining	13
1.3.2 Рекомендаційні системи.....	15
1.3.3 Контентная фільтрація	15
1.3.4 Колаборативна фільтрація (CF).....	16
1.3.5 Бассовський класифікатор	20
1.3.6 Матрична факторизація.....	23
1.4 Огляд аналогів рекомендаційних систем	26
1.5 Огляд інших систем, що надають картографічні дані та об'єкти відвідування.....	29
1.6 Оцінка результатів рішення задачі інтелектуального аналізу даних.....	31
2 ПРОЕКТУВАННЯ СИСТЕМИ РЕКОМЕНДАЦІЙ МІСЦЬ ВІДВІДУВАННЯ	33
2.1 Загальна архітектура системи.....	33
2.1.1 Глобальна конфігурація системи	33
2.1.2 Інструменти для обробки великих даних	35
2.1.3 Картографічні дані.....	37

2.1.4	Архітектура backend системи	38
2.2	Функціонально-компонентна архітектура системи	39
3	РЕАЛІЗАЦІЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ	41
3.1	Рішення для ресурсоємких фонових обчислювальних задач і процесів.....	42
3.2	Збір, зберігання і очищення навчальних перших даних об'єктів відвідування.....	49
3.3	Реалізація аналізу даних для отримання рекомендацій об'єктів відвідування.....	52
3.4	Оцінка отриманих рекомендацій.....	60
3.5	Можливі поліпшення системи.....	62
	ВИСНОВКИ.....	63
	СПИСОК ЛІТЕРАТУРИ.....	64
	ДОДАТОК А СХЕМИ БАЗИ ДАНИХ	67

СПИСОК СКОРОЧЕНЬ

ETL – Extract, transform, load;

TF – term frequency;

IDF – inverse document frequency;

CF – Collaborative Filtering;

SVD – Singular Value Decomposition;

ALS – Alternating Least Squares;

RMSE – root-mean-square error;

СУБД – система управління базами даних;

РСУБД – реляційна система управління базами даних;

БД – база даних;

ОС – операційна система;

RDD – Resilient Distributed Dataset;

HDFS – Hadoop Distributed File System;

OSM – Open Street Map;

MVC – model-view-controller;

DTO – Data Transfer Object;

СД – сервер додатків;

IoC – inversion of control;

ORM – object-relational mapping;

ВСТУП

З розвитком інформаційних технологій все більше різних областей діяльності використовують цифровий формат, надаючи нові можливості людям і покращуючи їх життя. Одним із значущих кроків в поліпшенні житті людей стала поява цифрових карт і агрегаторів картографічних даних з об'єктами для відвідування, навчання і відпочинку людей. У сучасному світі спосіб життя більшості людей складається з проведення часу в місцях харчування, освіти та розваги. Дана тенденція стимулює економіку і створює приємне дозвілля людям. Поява інструментів, що зберігають такі дані, значно поліпшило можливість людей знайти цікаве для них місце. Але на сьогоднішній день існують послуги тільки зберігання переліку об'єктів з їх місцем розташування, без будь-якого механізму якісного надання місць проведення часу, підлаштований під переваги користувача. Така можливість була б наступним кроком у розвитку такого важливого напрямку і поліпшення житті людей з сьогоднішніми тенденціями.

Побудова системи для надання рекомендацій в залежності від вподобань користувача вимагає застосування методів інтелектуального аналізу даних, а точніше методів Data Mining. Методи Data Mining пропонують рішення для інтелектуальної обробки великих масивів інформації з виявленням закономірностей в даних. Метою такої обробки служить побудова прогнозу або вироблення рекомендацій.

Рекомендаційні системи поєднують в собі всі механізми і рішення, необхідні для надання комплексу послуг з аналізом переваг великої кількості людей. Вподобаннями можна назвати явні і неявні ознаки закладів, які в певній мірі цікавлять окремих користувачів.

Актуальність роботи визначається тенденцією сучасного способу життя, яке передбачає відвідування людьми місць відпочинку, які їм найбільш цікаві.

Метою роботи є підвищення якості рекомендацій місць відпочинку на основі аналізу перевазі відпочиваючих.

Обрана тематика досліджень передбачає інтенсивну роботу з геоданими і необхідність зберігання великих обсягів даних за умови, що ці дані мають неоднозначний спосіб ідентифікації. У подібній системі неминуче сильне зростання обсягу інформації та необхідність у масштабуванні.

Для досягнення поставленої мети визначено наступне завдання: «Розробка системи, яка виступає в ролі сервісу, який на основі персональних даних про користувача надає функціонал для складання і видачі рекомендацій о місцях відпочинку, заснованих на інтелектуальному аналізі даних відгуків інших користувачів». Передбачається, що сам сервіс не наказує кінцевий контекст, в якому він буде використовуватися, що дозволяє застосування в будь-якому довільному середовищі. Для надання рекомендацій, необхідно розробити засоби для зберігання даних о користувачах, закладах, відгуках з інтерфейсом для комунікації з іншими системами-клієнтами, які будуть наповнювати рекомендаційну систему цими даними. Особливістю даної системи є накопичення відгуків кожного користувача на великому проміжку часу для тривалої побудови і поліпшення рекомендацій по місцям відпочинку в містах, де ці користувачі знаходяться.

Дане завдання розбито на декілька етапів:

- провести аналіз існуючих аналогів;
- провести аналіз можливих різних методів Data Mining для побудови рекомендаційних систем;
- розробити засоби для ефективного виконання ресурсоємних завдань;
- розробити засоби для отримання і зберігання картографічних даних;
- розробити інструменти для інтегрування навчальних даних з зовнішніх джерел з їх трансформацією і очищенням;

- інтегрувати функціонал для побудови моделі рекомендаційної системи а також реалізувати інтерфейси для видачі рекомендацій на її основі;
- провести оцінку ефективності моделі;

Об'єктом дослідження є модель надання рекомендацій місць відвідування, на основі вподобань користувачів.

Предметом дослідження є методи інтелектуального аналізу даних, що застосовуються в рекомендаційних системах.

ВИСНОВКИ

У даній роботі запропонована архітектура і модель побудови системи, що надає персоналізовані рекомендації по місцям відпочинку на основі аналізу відгуків про них. Проаналізовано множина різноманітних підходів для побудови моделі Data Mining і обраний метод, який надав хороше поєднання продуктивності і точності з обраним інструментарієм.

У дипломній роботі сформульовані вимоги до розроблюваної рекомендаційної системи. Виявлені і вирішені проблеми, що виникають при розробці подібних систем.

Виконано проектування системи з множиною компонент і складною комунікацією між ними. Проведена реалізація всіх компонент з необхідною для якісної розробки архітектурою. Розроблений механізм для виконання ресурсномістких завдань, на підставі якого додано множина багатокрокових процесів, необхідних для збору картографічних даних, збору даних для навчання, очищення, ефективного перетворення даних, навчання даних з комунікацією з кластером Spark і збереженням підрахованих рекомендацій для користувачів. В якості головного методу навчання моделі вирішено використовувати матричну факторизацію з оптимізацією ALS, заснованому на стохастичному спуску.

В якості результату отримана гнучка архітектура і модель з оцінкою RMSE 0,44 в діапазоні значень відгуків від 1 до 5.

СПИСОК ЛІТЕРАТУРИ

1. Han Jiawei. Data Mining, Concepts and Techniques – Waltham : Morgan Kaufmann, 2013. – 744 p.
2. Rajaraman Anand. Mining of Massive Datasets / Anand Rajaraman, Jeffrey Ullman – Cambridge : Cambridge University Press, 2011. – 326 p.
3. Murphy Kevin P. Machine Learning: A Probabilistic Perspective – London : The MIT Press, 2012. – 1104 p.
4. Falk Kim. Practical Recommender Systems – Shelter Island : Manning Publications, 2019. – 432 p.
5. Aditya P. H. A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for E-commerce in Indonesia: A case study PT X / P. H. Aditya, Indra Budi, Qorib Munajat // International Conference on Advanced Computer Science and Information Systems – Malang: IEEE, 2016. – P. 303-308.
6. Levinas Claudio Adrian. An Analysis of Memory Based Collaborative Filtering Recommender Systems with Improvement Proposals // International Conference on Advanced Computer Science and Information Systems – Bali, Indonesia : IEEE, 2017. – P. 120-127.
7. Meira Dania. An Experimental Analysis on Scalable Implementations of the Alternating Least Squares Algorithm / Dania Meira, Jose Viterbo, Flavia Bernardini // Proceedings of the Federated Conference on Computer Science and Information Systems – Gdansk, Poland : Polskie Towarzystwo Informatyczne, 2018. – P. 352-353.
8. Paul Covington, Jay Adams, Emre Sargin, Deep Neural Networks for YouTube Recommendations [Електронний ресурс] – Режим доступу:

- <https://static.googleusercontent.com/media/research.google.com/ru//pubs/archive/45530.pdf>.
9. Linden Greg. Amazon.com Recommendations Item-to-Item Collaborative Filtering / Greg Linden, Brent Smith, Jeremy York. // IEEE Internet Computing – Los Alamitos : IEEE, 2003. – P. 76-80.
 10. How Is Spotify's Thriving Recommendation System Becoming A New Advertising Platform? [Электронный ресурс] – Режим доступа: <https://medium.com/swlh/how-is-spotifys-thriving-recommendation-system-becoming-a-new-advertising-platform-a2b97ffe2012>.
 11. What are Mean Squared Error and Root Mean Squared Error? [Электронный ресурс] – Режим доступа: <https://medium.com/swlh/how-is-spotifys-thriving-recommendation-system-becoming-a-new-advertising-platform-a2b97ffe2012>.
 12. ML Tuning: model selection and hyperparameter tuning [Электронный ресурс] – Режим доступа: <https://spark.apache.org/docs/latest/ml-tuning.html>.
 13. Types of data in recommender systems [Электронный ресурс] – Режим доступа: http://manishbarnwal.com/blog/2018/09/27/types_data_recommender_system.
 14. Nishant Neeraj. Mastering Apache Cassandra – Birmingham : Packt Publishing, 2015. – 350 p.
 15. Hadoop or Spark? [Электронный ресурс] – Режим доступа: <https://phoenixnap.com/kb/hadoop-vs-spark>.
 16. How osm works [Электронный ресурс] – Режим доступа: <https://learnosm.org/en/osm-data/data-overview/#how-osm-works>.
 17. Key metrics for monitoring Tomcat [Электронный ресурс] – Режим доступа: <https://www.datadoghq.com/blog/tomcat-architecture-and-performance/>.
 18. Spring Beans in Depth [Электронный ресурс] – Режим доступа: <https://medium.com/javarevisited/spring-beans-in-depth-a6d8b31db8a1>.

19. Apache Mahout and Spark Comparison [Электронный ресурс] – Режим доступа: <http://matthewbarga.com/blog/index.php/2015/10/22/apache-mahout-and-spark-comparison/>.
20. Apache Spark Architecture Explained in Detail [Электронный ресурс] – Режим доступа: <https://www.dezyre.com/article/apache-spark-architecture-explained-in-detail/338>.