АНОТАЦІЯ

Міхав В. В. Модель та методи збору і обробки даних для рекомендаційних систем у peer-to-peer комп'ютерних мережах. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 123 Комп'ютерна інженерія. – Черкаський державний технологічний університет, Черкаси, 2023.

Актуальність дослідження зумовлена зростанням кількості неструктурованих даних у комп'ютерних мережах різних типів. Для покращення пошуку у великих масивах інформації все частіше звичайні методи пошуку доповнюють рекомендаційними системами. Рекомендаційні системи дозволяють полегшити пошук при великій кількості контенту, доповнюючи або заміняючи класичну пошукову видачу рекомендаціями. Зокрема, вони дозволяють полегшити пошук при великій кількості об'єктів у системі, доповнюючи класичну пошукову видачу рекомендаціями, а в деяких ситуаціях навіть заміняють пошук. Також рекомендаційні системи можуть застосовуватися для ранжування результатів класичного пошуку. Таким чином вони можуть різними способами поєднуватися із звичайними пошуковими алгоритмами. В P2P мережах застосування рекомендаційних систем може мати додаткову користь. Якщо користувач шукає конкретний файл, доданий до мережі раніше, і файл не знайдено з різних причин, можна надати користувачу список рекомендацій з врахуванням його вподобань і, можливо, пошукового запиту. У децентралізованих P2P мережах часто виникає проблема індексації та пошуку файлів на різних пристроях мережі.

P2P комп'ютерні мережі знову актуальні й широко використовуються у наш час завдяки численним перевагам, які вони пропонують. Якщо спочатку вони набули популярності завдяки файлообміним сервісам, то зараз їх використовують й у багатьох інших сферах. Найважливішою перевагою P2P мереж є децентралізація. Вони можуть працювати без центрального сервера,

що робить їх стійкішими до відмов та атак. Це особливо корисно в умовах збільшення кількості кіберзагроз. P2P також сприяють розвитку сучасних технологій, таких як блокчейн. Багато криптовалют та додатків для управління цифровими активами базуються на P2P принципах, що забезпечує їх безпеку та надійність. У сфері інтернет-телебачення, стрімінгу та онлайн комп'ютерних ігор P2P також стають все популярнішими. Вони дозволяють роздавати контент без великих централізованих серверів, що ефективно зменшує навантаження на інфраструктуру мережі та забезпечує якість стрімінгу. Тож P2P мережі залишаються актуальними завдяки своїй децентралізованій природі, яка сприяє безпеці та стійкості в умовах зростаючих кіберзагроз, а також вони допомагають у розвитку нових технологій та способів розповсюдження контенту.

З різних причин шукані файли можуть бути недоступні для користувача, навіть якщо вони були додані раніше до системи та проіндексовані. Наприклад, комп'ютери, що містять потрібний файл або таблиці маршрутизації до нього чи його частин, вийшли з мережі, або застосовуються технології побудови P2P мережі з ймовірнісними методами пошуку, що не завжди знаходять далеко розташовані від комп'ютера користувача файли тощо. В одноранговій комп'ютерній мережі вузли – учасники мережі динамічно під'єднуються та від'єднуються, при цьому можуть виникнути проблеми з достовірністю та безпекою даних інформаційної системи мережі, зокрема й рекомендаційної, адже нові вузли, що під'єднуються можуть управлятися зловмисниками або бути ураженими шкідливим програмним забезпеченням. Такі вузли можуть бути запрограмовані на викривлення або перехоплення даних рекомендаційної системи. Тому важливо при здійсненні пошуку і створенні рекомендацій враховувати також достовірність та інформаційну безпеку інформації. При формуванні пошукової видачі та створенні рекомендацій користувачам децентралізованих P2P комп'ютерних мереж також важливо враховувати час на доступ до інформації та знаходити баланс між точністю та швидкістю

отримання результатів. Оскільки інформація та таблиці маршрутизації в децентралізованих мережах зберігаються розподілено, важливо враховувати, що деякі файли користувач може отримати швидше, ніж інші, а також, що інколи час пошуку і завантаження файлу може зробити неактуальним зусилля на отримання доступу до нього. Важливим питанням є також кількість пам'яті, яку можна виділити під дані рекомендаційної системи, адже у децентралізованій P2P мережах їх доведеться зберігати розподілено на комп'ютерах учасників мережі. Вибір методу представлення даних, якими оперує рекомендаційна система, має вагомий вплив, оскільки ефективний спосіб представлення даних, необхідних для роботи такої системи, може зменшити кількість потрібних ресурсів та збільшити кількість доступних алгоритмів для формування списків рекомендацій.

Тож рекомендаційні системи значним чином впливають на те, яким користувачі сприймають інформаційний простір, а їх інформаційна безпека та часова і просторова складність алгоритмів є важливими складниками забезпечення якості та ефективності роботи систем пошуку і фільтрації даних у комп'ютерних мережах, особливо децентралізованих.

На основі проведеного дослідження можливих шляхів вирішення наявної науково-практичної задачі було наступним чином сформульовано мету дисертаційної роботи: зменшення витрат часу та пам'яті на роботу рекомендаційної системи в однорангових децентралізованих комп'ютерних мережах при достатній точності створення рекомендацій користувачам.

Проведено дослідження та порівняльний аналіз методів роботи peer-to-peer комп'ютерних мереж, методів зберігання та пошуку даних у них, а також моделей децентралізованих рекомендаційних мереж.

Вперше розроблено математичну модель процесів збору і обробки даних для рекомендаційної системи в одноранговій децентралізованій комп'ютерній мережі, яка відрізняється від відомих можливістю оцінки ймовірнісно-часових характеристик процесів формування і зміни рекомендацій за допомогою мультитригерного GERT-моделювання та

врахуванням вимог достовірності і безпеки даних під час змін у структурі мережі, що дозволяє здійснювати раціональний вибір параметрів системи.

Удосконалено метод зберігання даних рекомендаційної системи, який відрізняється від відомих адаптацією до архітектури однорангових децентралізованих комп'ютерних мереж та використанням хеш-таблиць для зберігання даних про користувачів і об'єкти контенту та зв'язних списків для зберігання даних про рекомендації, що дозволило зменшити витрати часу і пам'яті на процеси обробки даних системи.

Удосконалено метод пошуку та фільтрації даних рекомендаційною системою для формування рекомендацій користувачам, який відрізняється від відомих використанням запропонованого методу зберігання даних для збереження проміжних та підсумкових результатів обчислень, що дозволило використовувати його в однорангових децентралізованих комп'ютерних мережах та зменшити витрати часу і пам'яті при забезпеченні високої точності прогнозування вподобань користувачів.

Проведена оцінка якості та ефективності запропонованих моделі та методів шляхом проведення експериментів на програмній імітаційній моделі, а також експериментів із використанням відкритих наборів даних Netflix Prize data та MovieLens.

Практична цінність роботи полягає у такому:

– Розроблена імітаційна модель процесів збору та обробки даних для рекомендаційної системи, що дозволила оцінити ймовірнісно-технічні характеристики рекомендаційної системи децентралізованої однорангової комп'ютерної мережі. Отримано аналітичний вираз, за допомогою якого є можливість оцінити щільність розподілу ймовірностей часу ідентифікації стану вузлів децентралізованої рекомендаційної системи. Використання мультитригерного підходу та врахування вимог до достовірності та безпеки рекомендаційних повідомлень в порівнянні з відомими моделями дозволило підвищити точність результатів моделювання до 5%. Розроблено алгоритми моделювання рекомендаційної системи однорангової децентралізованої

комп’ютерної мережі та процесів у ній, що дають можливість проводити тестування різних методів зберігання, пошуку та фільтрації даних у мережі.

– Розроблено алгоритми зберігання даних рекомендаційної системи однорангової децентралізованої комп’ютерної мережі, що дозволяють зменшити витрати пам’яті і часу на процеси зберігання та читання даних системи. Зокрема, розроблена система на основі запропонованих алгоритмів при використанні у децентралізованій одноранговій мережі показує наступні результати в порівнянні з найкращими результатами відомих систем – в 2,4 разів кращі результати по часу заповнення бази даних, в 1,7 разів кращі результати по використаному об’єму пам’яті та в 2,5 разів кращі результати по часу генерації рекомендацій.

– Розроблено алгоритм пошуку та фільтрації даних рекомендаційною системою для формування рекомендацій користувачам в однорангових децентралізованих комп’ютерних мережах, що зменшує витрати часу і пам’яті при забезпеченні високої точності прогнозування вподобань користувачів. Точність розробленого алгоритму сягає до 0,84 в залежності від обраних параметрів системи.

Практичне значення отриманих результатів підтверджено відповідними актами впровадження. Результати дисертаційних досліджень впроваджені і використовуються у діяльності ІТ-компанії ТОВ "ОНІКС-СИСТЕМЗ", а також використано у навчальному процесі Центральноукраїнського національного технічного університету.

Ключові слова: комп’ютерні мережі, рекомендаційні системи, GERT-мережі, структури даних, бази даних, програмне імітаційне моделювання, децентралізовані однорангові комп’ютерні мережі, бінарні діаграми рішень, дерева рішень, зв’язні списки, хеш-таблиці, фільтрація даних, пошук даних.

SUMMARY

Mikhav V. V. Model and methods of data receiving and processing for recommendation systems in peer-to-peer computer networks. – Qualifying scientific work on the rights of the manuscript.

Dissertation for the degree of Doctor of Philosophy (PhD) in the specialty 123 “Computer Engineering”. – Cherkasy State Technological University, Cherkasy, 2023.

The relevance of the research is driven by the increasing volume of unstructured data in various types of computer networks. To enhance search capabilities within vast information repositories, conventional search methods are increasingly complemented by recommendation systems. Recommender systems make it easier to search in a large amount of content by supplementing or even replacing traditional search results with recommendations. In particular, they ease the search process when dealing with an abundance of items within a system supplementing the classic search output with recommendations, and in some cases, even substituting search altogether. Recommendation systems can also be applied to rank classical search results, thus providing various ways to integrate with conventional search algorithms. In P2P networks, the application of recommendation systems can offer additional benefits. If a user is searching for a specific file that was previously shared on the network but cannot be located for various reasons, recommendations can be provided to the user, taking into account their preferences and, perhaps, their search query. In decentralized P2P networks, the problem of indexing and locating files on various network devices often arises.

P2P networks have regained relevance and widespread usage due to their multiple benefits. Though these networks originally gained popularity via file-sharing services, they now serve many other purposes. Decentralization is P2P networks' most important advantage. They can function without a central server, rendering them more resilient to failures and attacks, especially in the current era of escalating cyber threats. P2P networks also aid in the advancement of contemporary

technologies such as blockchain. Numerous cryptocurrencies and digital asset management programs rely on P2P principles, guaranteeing their safety and dependability. In the realm of Internet television, streaming, and online computer games, P2P services are also growing in popularity. Peer-to-peer (P2P) networks enable content distribution without relying on large centralized servers, effectively reducing network infrastructure load and enhancing the quality of streaming. P2P networks remain pertinent due to their decentralized nature, which enhances security and resilience against increasing cyber threats, and contributes to advancing new technologies and content distribution methods.

Files that have been added to the system and indexed may still not be accessible to the user due to various factors. For instance, the computers containing the required file or routing tables may have gone offline, or probabilistic search methods utilized in P2P network technologies may fail to locate files distant from the user's computer. In a peer-to-peer computer network, the nodes of network participants establish and terminate connections dynamically. This process may degrade the reliability and security of network information system data, including the recommendation system. Such problems can occur due to newly connected nodes that are controlled by attackers or affected by malware. These nodes may be programmed to interfere with the data or even cause corruption of the recommendation system. Therefore, it is important to prioritize reliability and information security when searching for and recommending information.

When generating search results and recommendations for users of decentralized P2P computer networks, it is also important to consider the time it takes to access information and strike a balance between accuracy and speed of results. Because information and routing tables are stored in a distributed manner on decentralized networks, it is important to consider that some files can be retrieved faster than others, and that sometimes the time it takes to find and download a file can make access to it irrelevant. Another important issue is the amount of memory that can be allocated to the recommendation system data, since in a decentralized P2P network it will have to be stored distributed on the computers of the network

participants. The choice of how to represent the data used by a recommendation system has a significant impact, since an efficient way of representing the data required by the system can reduce the amount of resources required and increase the number of algorithms available for generating recommendation lists.

Therefore, recommendation systems have a significant impact on how users perceive the information space, and their information security as well as the temporal and spatial complexity of algorithms are important components of ensuring the quality and efficiency of data search and filtering systems in computer networks, especially decentralized ones.

Based on the conducted research on possible ways to address the existing scientific and practical problem, the aim of the dissertation was formulated as follows: to reduce the time and memory costs of operating a recommendation system in peer-to-peer decentralized computer networks while ensuring sufficient accuracy in generating recommendations for users.

A study and comparative analysis of peer-to-peer computer network operation methods, data storage and retrieval methods within them, as well as models of decentralized recommendation networks, have been conducted.

For the first time, a mathematical model of data collection and processing processes for a recommender system in a peer-to-peer decentralized computer network has been developed. It differs from the known ones by the possibility of assessing the probabilistic and temporal characteristics of the processes of forming and changing recommendations using multi-trigger GERT modeling and taking into account the requirements for data reliability and security during changes in the network structure, which allows to make a rational choice of system parameters.

The method of storing data of the recommender system has been improved, which differs from the known ones by adapting to the architecture of peer-to-peer decentralized computer networks and using hash tables to store data on users and content objects and linked lists to store data on recommendations, which reduced the time and memory consumption for data processing of the system.

The method of searching and filtering data by a recommender system for

generating recommendations to users has been improved. It differs from the known ones by using the proposed method of data storage to store intermediate and final results of calculations, which made it possible to use it in peer-to-peer decentralized computer networks and reduce time and memory costs while ensuring high accuracy of predicting user preferences.

The quality and effectiveness of the proposed model and methods were evaluated by conducting experiments on a software simulation model, as well as experiments using the open datasets Netflix Prize data and MovieLens.

The practical value of the work lies in the following:

– A simulation model of data collection and processing processes for a recommender system has been developed, which has made it possible to evaluate the probabilistic and technical characteristics of a recommender system of a decentralized peer-to-peer computer network. An analytical expression has been obtained that makes it possible to estimate the probability density function of the time of identifying the state of the nodes of a decentralized recommender system. The use of a multi-trigger approach and consideration of the requirements for the reliability and security of recommendation messages in comparison with known models allowed to increase the accuracy of the modeling results by up to 5%. The algorithms for modeling the recommender system of a peer-to-peer decentralized computer network and processes in it have been developed, which make it possible to test various methods of storing, searching and filtering data in the network.

– The algorithms for storing data of the recommender system of a peer-to-peer decentralized computer network have been developed, which allow reducing the memory and time consumption for storing and reading system data. In particular, the developed system based on the proposed algorithms, when used in a decentralized peer-to-peer network, shows the following results compared to the best results of known systems: 2.4 times better results in terms of database filling time, 1.7 times better results in terms of memory usage, and 2.5 times better results in terms of recommendation generation time.

– An algorithm for searching and filtering data by a recommender system

for generating recommendations to users in peer-to-peer decentralized computer networks has been developed, which reduces time and memory costs while ensuring high accuracy in predicting user preferences. The accuracy of the developed algorithm reaches up to 0.84, depending on the selected system parameters.

The practical significance of the results obtained is confirmed by the relevant acts of implementation. The results of the dissertation research have been implemented and are utilized in the activities of the IT company "ONIX-SYSTEMS" LLC and have also been incorporated into the educational process at the Central Ukrainian National Technical University.

Keywords: computer networks, recommender systems, GERT networks, data structures, databases, computer simulation modeling, decentralized peer-to-peer computer networks, binary decision diagrams, decision trees, linked lists, hash tables, data filtering, data search.