Технічні науки

УДК 004.8:004.6:615

**Бовсуновська Катерина Сергіївна**

*старший викладач кафедри біомедичної кібернетики*

*Національний технічний університет України*

*«Київський політехнічний інститут імені Ігоря Сікорського»*

**Bovsunovska Kateryna**

*Senior Lecturer at the Biomedical Cybernetics Department*

*National Technical University of Ukraine*

*"Igor Sikorsky Kyiv Polytechnic Institute"*

**Кравець Олексій Володимирович**

*магістрант*

*Національного технічного університету України*

*«Київський політехнічний інститут імені Ігоря Сікорського»*

**Kravets Oleksii**

*Master's Student of the*

*National Technical University of Ukraine*

*"Igor Sikorsky Kyiv Polytechnic Institute"*

# СИСТЕМА АНАЛІЗУ ВЕЛИКИХ БАЗ ДАНИХ З ВИКОРИСТАННЯМ ДЕРЕВ РОЗВ'ЯЗКІВ

# SYSTEM FOR ANALYZING LARGE DATABASES USING DECISION TREES

***Анотація.*** *Вступ. Зростання потужності комп'ютерів та інтернету призвело до значного збільшення обсягів даних, що збираються організаціями та особами, і цей тренд лише зростає. Таке зростання впливає на багато наук, змінюючи підходи в медицині, біології, соціології та*

*інших галузях. Сучасні технічні розвитки дозволяють застосовувати складні алгоритми, штучні нейронні мережі, та імітаційні моделі для розв'язання задач, які раніше вимагали експертного підходу. Зростання обчислювальних потужностей також дозволило використовувати алгоритми, що раніше були обмежені науковим інтересом, для реальних багатовимірних завдань.*

*Великі дані все більше проникають у малий і середній бізнес, не тільки для аналізу власної статистики, але й для використання даних великих операторів в аналітичних цілях. При цьому підприємці стикаються з високими витратами на використання таких даних, відсутністю кваліфікованих спеціалістів та необхідних обчислювальних потужностей.*

*Метою дослідження є розробка концепції, моделювання, конструювання та програмна реалізація спеціалізованої рекомендаційної системи, що використовує гібридну модель. Джерелом даних для системи є експертні знання та додаткова інформація про лікарські засоби, включаючи склад, аналоги, протоколи лікування та оцінки їх успішності. Проект надає підприємствам можливість доступу до високоякісних аналітичних інструментів, мінімізуючи потребу у значних інвестиціях у розвиток власних аналітичних відділів.*

*Методами та інструментами дослідження є машинне навчання та штучні нейронні мережі; метод ранжування, що базується на сесіях взаємодії з користувачем та алгоритмі RankBoost; метод градієнтного бустінгу та його реалізації XGBoost; моделювання та конструювання програмного забезпечення в нотації UML.*

*Результати. В роботі розроблено концепцію, сконструйовано та реалізовано програмно запропоновану модель рекомендаційної системи. Після більш глибокого тестування та налаштування, а також розширення інформаційного базису системи, вона може бути застосована як комерційний проект.*

*Перспективи. В подальших дослідження пропонується зосередити увагу на розширені інформаційної бази функціонування системи та збільшенні кількості функцій. Також доцільним є впровадження метрик оцінки якості рекомендацій, що можуть бути використані в алгоритмі самостійного додаткового навчання алгоритмів системи та при тонкому налаштуванні параметрів цих алгоритмів.*

***Ключові слова:*** *рекомендаційна система, ранжування, сесія користувача, хмарна система, експертний метод, гібридна модель.*

***Summary.*** *Introduction. The growth in computer power and the internet has led to a significant increase in the volumes of data collected by organizations and individuals, and this trend is only growing. This growth affects many sciences, changing approaches in medicine, biology, sociology, and other fields. Modern technical developments allow the use of complex algorithms, artificial neural networks, and simulation models to solve tasks that previously required an expert approach. The increase in computational power has also enabled the use of algorithms that were previously limited to scientific interest for real multidimensional tasks.*

*Big data is increasingly penetrating small and medium-sized businesses, not only for analyzing their own statistics but also for using the data of large operators for analytical purposes. In this case, entrepreneurs face high costs of using such data, a lack of qualified specialists, and necessary computational capacities.*

*Purpose is to develop a concept, model, construct, and implement a specialized recommendation system that uses a hybrid model. The source of data for the system includes expert knowledge and additional information about medical substances, including composition, analogs, treatment protocols, and their success rates. The project provides businesses with access to high-quality*

*analytical tools, minimizing the need for significant investments in developing their own analytical departments.*

*Materials and methods include machine learning and artificial neural networks; a ranking method based on user interaction sessions and the RankBoost algorithm; the gradient boosting method and its XGBoost implementation; modeling and constructing software in UML notation.*

*Results. The paper develops a concept, constructs, and implements the proposed model of the recommendation system. After more extensive testing and calibration, as well as expanding the system's information base, it can be applied as a commercial project.*

*Discussion. Further research proposes to focus on expanding the information base of the system's operation and increasing the number of functions. It is also advisable to introduce metrics for evaluating the quality of recommendations, which can be used in the algorithm for self-learning of the system's algorithms and for fine-tuning the parameters of these algorithms.*

***Key words:*** *recommendation system, ranking, user session, cloud system, expert method, hybrid model.*

**Постановка проблеми.** Рекомендаційні системи, один з підрозділів машинного навчання [1], є алгоритмами, що підбирають релевантні товари, послуги та контент на базі даних про ці сутності та користувачів. Зазвичай рекомендаційні системи потребують великих обчислювальних потужностей разом з засобами накопичення, впорядкування та аналізу великої кількості даних, що збираються у потоковому режимі. Це суттєво обмежує можливості використанням такого потужного маркетингового інструменту малим й та середніми бізнесами, що в більшості не мають відповідних ресурсів. Роботою запропоновано наданням подібним бізнес – клієнтам можливості користування такою системою на базі підписки. Система реалізується у вигляді API [2], що узгоджується з торговими системами

клієнтів, та обмінюється з ними необхідною для генерації рекомендацій інформацією. Джерелом даних для системи є експертні знання та додаткова інформація про лікарські засоби, включаючи склад, аналоги, протоколи лікування та оцінки їх успішності.

**Метою статті** є розкриття концептуальних підходів до створення такої системи, у тому числі адаптації математичних моделей та алгоритмів, використовуваних рекомендаційними системами, до предметної галузі – медичних препаратів. Також стаття демонструє ключові етапи побудови та реалізації системи на базі сучасних технологій, технік та інструментів.

**Матеріали і методи.** Матеріалами дослідження є праці зарубіжних авторів, що провадять свої наукові дослідження у галузі алгоритмів машинного навчання, зокрема рекомендаційних систем та дерев рішень, а також сучасні керівництва з проектування, реалізації та розгортання розподілених систем з клієнт – серверною архітектурою.

В процесі дослідження було використано наукові методи групування, аналізу та синтезу для побудови аналітико – математичного базису створеної системи. При конструюванні системи застосовано нотацію UML, що дозволила проаналізувати та узагальнити усі сценарії використання системи, формалізувати ключові алгоритми, структурувати джерела даних та програмні компоненти, надати уявлення про архітектуру та розгортання системи на усіх необхідних рівнях абстракції. При тестування системи послідовно використані методи модульного, функціонального та A/B тестування. При створенні стартап – проекту застосовувались економічні підходи, а формулювання висновків є результатом застосування методу логічного узагальнення.

**Виклад основного матеріалу.** Створена рекомендаційна система використовує гібридну модель [1], що дозволяє з одного боку уникнути типових недоліків систем, що базуються лише на одному алгоритмі, як от «проблема холодного старту», ризики «фільтраційних бульбашок», слабка

адаптація до зміни інтересів та фокусу користувача. З іншого боку, така побудова системи дозволяє врахувати експертні знання про протоколи лікування захворювань, статистику успішності лікування, а також властивості самих препаратів, у тому числі інформацію щодо діючих речовин та замінників.

Модель взаємодії з користувачами системи спроектована так, щоб ефективно розрізняти та класифікувати [3] різні типи потреб користувачів. Вона дозволяє відрізняти тимчасові, такі як одноразова потреба в назальних краплях, від регулярних, як-от періодичне використання певних вітамінів чи стимуляторів. Також модель здатна ідентифікувати постійні потреби, наприклад, у препаратах для лікування гіпотиреозу, які необхідно приймати щоденно і, відповідно, регулярно купувати. Базисом моделі є механіка сесій взаємодії з користувачем [1] та дерева рішень [4].

З практики відомо, що протоколи лікування більшості хвороб включають декілька препаратів: до антибіотиків призначують пробіотики, при лікуванні ОРВІ призначають препарати від нежиті, кашлю, антигістамінні та вітаміни, лікування опіків потребує не тільки знеболювальних, але й прискорювачів загоєння та знов таки антибіотиків, метою застосування яких є уникнення інфікування та ускладнень. Класифікатор, що побудований на базі дерев рішень та використовує експертні знання [4, 5], дозволяє обрати короткий перелік можливих хвороб та інших медичних кейсів, базуючись на історії інтересу користувача в рамках сесії та протоколах лікування. Саме цей алгоритм дозволяє згенерувати коректну пропозицію тих препаратів, що входять у той самий протокол лікування, що й препарати переглянуті та обрані користувачем. Накопичення історії сесій та аналіз даних як часових рядів [5], дозволяє додати до функціональності системи пропозиції препаратів, в яких користувач має підтверджену періодичну потребу.

Для збільшення зручності користування для B2B користувачів, функціонал системи розширено можливостями надання інформації про препарати, діючи речовини, аналоги, препаративні форми, тощо. Також окремий бізнес – користувач може переглядати звіти про генерації рекомендацій за довільний визначений період. Схематично можливості системи проілюстровані через структуру ендпоінтів REST API на рис. 1.

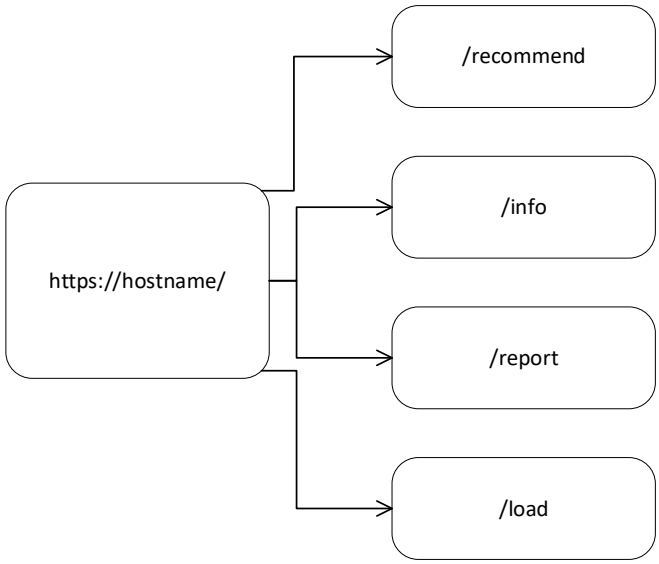


**Рис. 1. Структура ендпоінтів програмного інтерфейсу**

/recommend – відповідає за створення рекомендацій згідно надісланих в запиті даних. Тип запиту – GET.

/info – відповідає за відгук з переліком інформації згідно надісланих в запиті даних. Тип запиту – GET.

/report – відповідає за відгук із запитаним звітом, одним з сформованого в попередньому розділі переліку. Тип запиту – GET.

/load – завантаження оновленої інформації до системи. Тип запиту – POST.

Для тестування систему було розгорнуто відповідно до діаграми, що показана на рис. 2. На діаграмі окрім серверної частини позначено web-додаток, що був створений саме для тестування системи. Він спрощено імітує інтерфейс web-додатку, що може бути запропонований кінцевому споживачу B2B користувачем створеного порталу.

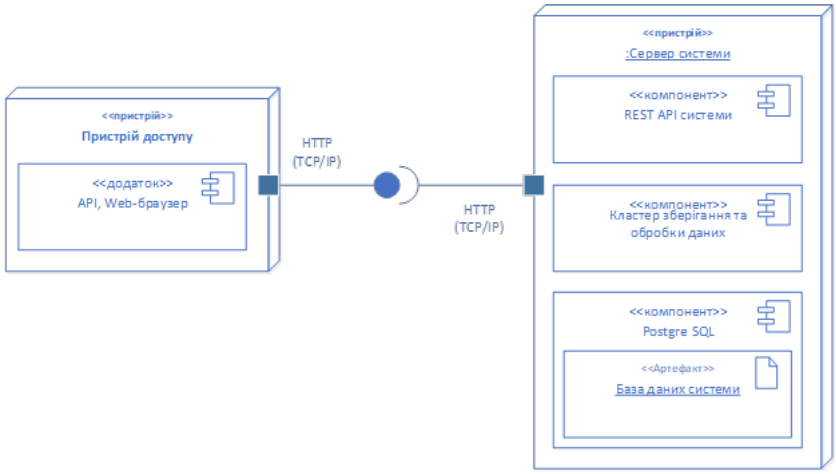


**Рис. 2. Діаграма розгортання**

Для розгортання системи використано контейнеризацію та оркестрацію під управлінням Docker. У разі необхідності масштабування системи можливо розгортання контейнерів на окремих серверах, що дозволяє масштабування тільки тих компонентів, що необхідно.

Тестування системи було організовано в два етапи окрім модульного тестування: запити до ендпоінтів із застосуванням Postman та користувацьке A/B тестування на базі створеного демонстраційного додатку.

Демонстраційний інтерфейс складається з однієї сторінки пошуку лікарських засобів, на якій також відображаються рекомендації системи, побудовані на історії взаємодії з користувачем. Інтерфейс має опції реєстрації та входу користувача, без яких накопичення історії взаємодій неможливе. Після авторизації користувача генерується JW токен, що є ідентифікатором сесії. Обмежений час придатності токену підвищує безпеку системи.

Після авторизації вводиться назва препарату в рядок пошуку. Розглянемо послідовність дій та результат генерації рекомендацій на прикладі препарату «Цефікс», що відтворено на рис. 3.

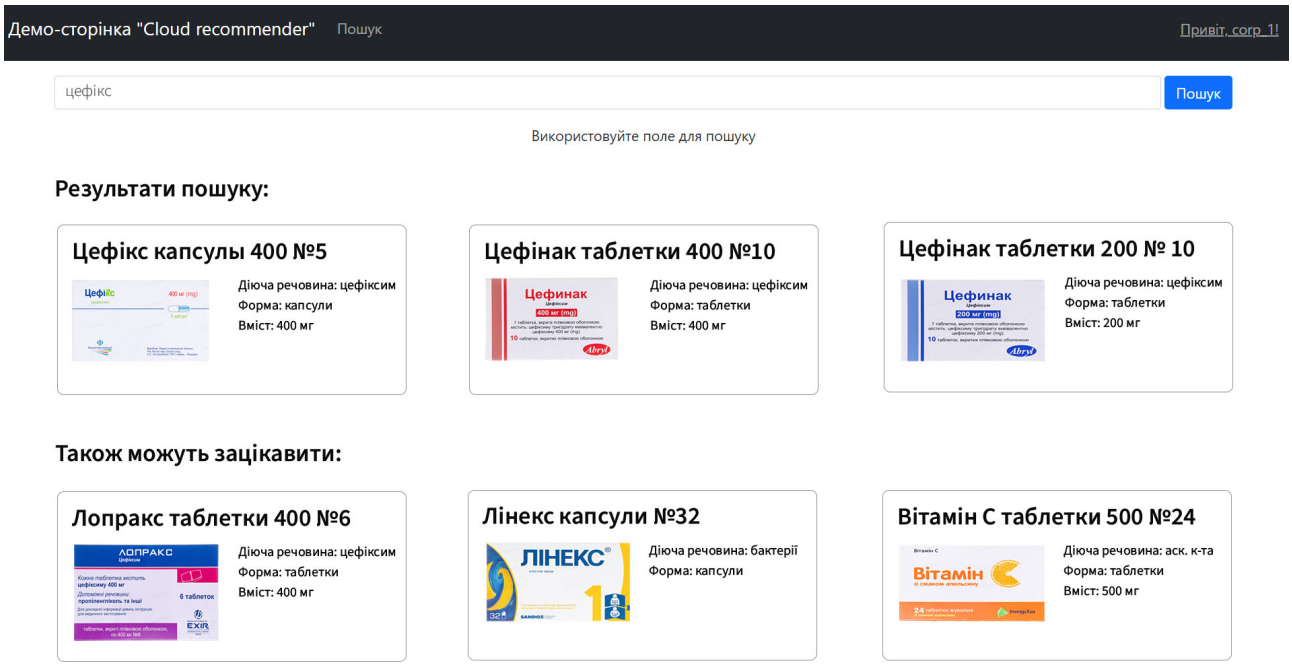


**Рис. 3. Пошук та видача рекомендацій**

В першому рядку карток (рис. 3), відображаються знайдені медичні препарати, а в другому – згенеровані рекомендації. Розберемо другий рядок більш детально:

«Лопракс» - фармакологічний аналог за діючою речовиною;

«Лінекс» - пробіотик, що призначають разом з попереднім препаратом, та препаратом, відображеним в результатах пошуку;

«Вітамін С» - препарат, що призначають у багатьох схемах лікування застудних захворювань.

**Висновки і перспективи подальших досліджень.** Розробка хмарної рекомендаційної системи, що взаємодіє з CRM-системами підприємств через API, є відповіддю на потреби ринку, де малі та середні бізнеси прагнуть використовувати сучасні маркетингові інструменти для збільшення продажів. Виконано повний цикл розробки концепції та прототипу, а також програмну реалізацію. Розроблена система базується на інноваційній гібридній моделі, що дозволила уникнути типових проблем рекомендаційних систем, актуальних моделях машинного навчання та

сучасній мікросервісній архітектурі. Економічне обґрунтування у вигляді розрахунку стартап проекту доводить економічну доцільність подальшого розвитку та вдосконалення створеного прототипу системи та залучення до співпраці користувачів B2B сегменту.

В подальших дослідження пропонується зосередити увагу на розширені інформаційної бази функціонування системи та збільшенні кількості функцій. Також доцільним є впровадження метрик оцінки якості рекомендацій, що можуть бути використані в алгоритмі самостійного додаткового навчання алгоритмів системи та при тонкому налаштуванні параметрів цих алгоритмів.

## Література


1. Croft W., Metzler D., Strohman T. Search engines: Information retrieval in practice. *Addison-Wesley.* 2010. 552 p.
2. Richardson L., Amundsen M. RESTful Web APIs: Services for a Changing World. *O'Reilly media.* 2013. 404 p.
3. Tianqi C., Guestrin C. XGBoost: A Scalable Tree Boosting System. *22nd ACM SIGKDD International Conference,* 2016. doi: https://doi.org/10.1145/2939672.2939785
4. Song T., Hu T. Research on XGboost academic forecasting and analysis modelling. *Second International Conference on Physics, Mathematics and Statistics.* Hangzhou, China, 2019. doi: 10.1088/1742-6596/1324/1/012091.
5. Jones H. An Essential Beginners Guide to Artificial Neural Networks and Their Role in Machine Learning and Artificial Intelligence. *CreateSpace Independent Publishing Platform.* 2018. 76 p.