MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică şi Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susţinere
Şef departament:
FIODOROV Ion dr., conf.univ.

„___” _____ 2024

EVALUAREA STATISTICĂ A SISTEMELOR DE RECOMANDARE ÎN STREAMING ȘI MAGAZINE ONLINE

Proiect de master

Student: _____ Balaur Dorina, TI-221M

Coordonator: _____ Cojocaru Svetlana, asist.univ.

Consultant: _____ Cojocaru Svetlana, asist.univ.

Chişinău, 2024

REZUMAT

Lucrarea vizează domeniul tehnologiei informației, în special explorarea domeniului recomandărilor pentru o experiență online al unui utilizator. Se vor cerceta sistemele de recomandare care stau la baza a mai multor platforme online și modelele care pot fi utilizate în crearea acestora. Se va determina importanța existenței recomandărilor atât pentru un utilizator simplu, cât și impactul care îl poate juca un sistem de recomandare asupra unei afaceri. La fel, se va cerceta necesitatea personalizării recomandărilor, dar și pericolele ce pot apărea în cazul când sistemul devine prea bun în a cunoaște utilizatorul și preferințele acestuia.

Pentru un sistem de recomandare, este important ca atât utilizatorul să se simtă valoros prin recomandările care i se fac, economisind timp din a căuta un timp îndelungat un produs, dar și pentru o afacere online este important de a folosi informațiile despre preferințele utilizatorilor în mod eficient. Cu acest scop există o mulțime de metrici, care, în dependență de tipul de platformă, setul de date și specificul afacerii, pot fi utilizați pentru a evalua eficiența sistemului de recomandare. Se vor analiza unii dintre cei mai populari metrici și se va cerceta cum și în ce context pot fi aplicați. La fel, în dependență de setul de date disponibil se vor analiza tipurile de metode de evaluare ale acestora utilizând metricii. Baza unui sistem de recomandare fiind modelul acestuia, se vor cerceta diferite tipuri de modele matematice existente și specificul fiecăruia.

Pentru partea practică a lucrării a fost utilizat *Jupyter Notebook* și limbajul de programare *Python.* Aceste intrumente sunt destul pentru a gestiona, procesa și analiza volume de date, ceea ce a permis explorarea într-un mod interactiv a două seturi de date cu care s-a lucrat ulterior.

Lucrarea analizează 2 seturi de date: unul cu *feedback* implicit și cu altul *feedback* explicit, pentru a putea forma o opinie despre provocările fiecăruia. Fiecare set va trece printr-o analiză exploratorie pentru a identifica specificul acestora. În dependență de set, se vor analiza modelele care ar putea fi aplicate în crearea unui sistem de recomandare. Spre final, sistemele create vor fi evaluate statistic și, unde posibil, se vor adăuga optimizări pentru a primi un răspuns al eficienței cât mai bun.

ABSTRACT

The paper focuses on the field of information technology, specifically exploring the field of content recommendations for a user's online experience. Will be explored multiple recommendation systems used on several online big known platforms and the models that are used at the base for their creation. A focus will be on determining the importance of a recommendation to both a simple user and the impact one recommandation can have on a business, depending if it was an inspired one or not. Similarily, the necessity of personalizing the recommendations will be researched, but also the dangers that may arise if the system becomes too good at knowing the user and predicting his preferences.

For a recommendation system, it is important that the user feels valuable through the recommendations that are made to him, saving time from searching for a long time for a product, but also for an online business it is important to use information about user preferences in an efficient way. Because of this there are a lot of metrics, which, depending on the type of platform, the dataset and the specifics of the business, can be used to evaluate the effectiveness of the recommendation system. The paper will investigate some of the most popular metrics and explore how and in what context they can be applied. Likewise, depending on the available dataset, different types of evaluation methods using the researched metrics will be analyzed. Considering that at basis of a recommendation system exists a model, different types of existing mathematical models and the specifics of each will be researched.

For the practical part of the work, *Jupyter Notebook* and the *Python* programming language were used. These tools are enough to manage, process and analyze volumes of data, which allowed the exploration in an interactive way of two data sets that were worked with later.

The paper analyzes 2 data sets: one with implicit feedback and another with explicit feedback, in order to form an opinion about the challenges of each. Each set will go through an exploratory analysis to identify their specifics. Depending on the data set, models that could be applied in creating a recommender system will be analyzed. Towards the end, the created systems will be statistically evaluated and, where possible, optimizations will be added to get the best efficiency response.

CUPRINS

INTRODUCERE

În prezent, considerând că o bună majoritate a oamenilor au acces la internet de pe cel puțin un dispozitiv, este logic să presupunem că un utilizator a efectuat cel puțin o dată o căutare de produs online. Aproape orice afacere care există în mod fizic are nevoie și de o prezență online pentru a putea primi vizibilitate și a fi considerată de către clienți. Considerând că bazele de date cu produse pot varia considerabil în dimensiune, este necesar ca să existe un sistem la bază care să recomande utilizatorului anume produsele de care acesta are nevoie într-un timp scurt. Aici sunt incluse sistemele de recomandare bazate pe filtrare.

Lucrarea dată va cerceta sistemele de recomandare existente pentru două seturi de date din domenii diferite precum platforme de *streaming* și magazin online. Se va încerca o cercetare a tehnicilor existente care stau la baza creării unui astfel de sistem și identificarea elementelor necesare pentru a construi un sistem de recomandare de succes. Prin cercetarea avantajelor și dezavatajelor utilizării unui astfel de sistem, se propune a identifica cum acesta influențează deciziile zilnice ale utilizatorilor, dar și modurile în care un asemenea sistem impactează o afacere online.

Tema lucrării se axează pe analiza sistemelor de recomandare și evaluarea statistică a eficienței acestora. Fiecare sistem informațional are un scop anume în dependență de tipul acestuia și natura afacerii, astfel scopul unui sistem de recomandare poate fi diferit. Chiar dacă accentul ar trebui să fie pe atragerea consumatorului pentru o perioadă mai lungă de timp pentru a genera profit, o astfel de strategie poate la fel de bine să fie negativă.

Un sistem de recomandare joacă un rol important în imaginea unui produs online, astfel că o recomandare bună sau o recomandare rea mereu va juca un rol în experiența unui utilizator. Pentru acest lucru există diverse modele care stau la baza sistemului. Lucrarea va cerceta și analiza tipurile de modele care pot fi folosite în crearea unui sistem de recomandare. Considerînd că nu există un model perfect, iar dacă acesta ar exista atunci ar fi o problema de supraadaptare, fiecare iterație de model trebuie analizată și înbunătățită. Lucrarea va evalua statistic modelul din punct de vedere a eficienței și se vor identifica punctele unde trebuie atrasă atenția.

Se va cerceta domeniul tehnologiei informaționale în combinație cu evoluția strategiilor de marketing digital, având ca scop final personalizarea conținutului recomandărilor distribuite consumatorilor de conținut online. Utilizând două seturi de date, se va încerca o analiză exploratorie a datelor, crearea sistemelor de recomadare potrivit seturilor de date și tipului de feedback. Ulterior, după înbunătățiri și iterații se va evalua eficiența sistemelor create și se vor formula concluzii.

BIBLIOGRAFIE

[1] „How Many Ads Do We See a Day? 17 Insightful Stats from 2023”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://webtribunal.net/blog/how-many-ads-do-we-see-a-day/

[2] „The Ultimate Guide to Brand Awareness”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://blog.hubspot.com/marketing/brand-awareness

[3] „Internet and social media users in the world 2023”, Statista. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.statista.com/statistics/617136/digital-population-worldwide/

[4] „Filter Bubbles as the Reality of the Internet”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.eurasian-research.org/publication/filter-bubbles-as-the-reality-of-the-internet/

[5] D. Banerjee, „Recommender Systems: An Overview of the Mathematics”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.opensourceforu.com/2022/07/recommender-systems-an-overview-of-the-mathematics/

[6] „Digital Media Literacy: How Filter Bubbles Isolate You”, GCFGlobal.org. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://edu.gcfglobal.org/en/digital-media-literacy/how-filter-bubbles-isolate-you/1/

[7] U. Hanani, B. Shapira, și P. Shoval, „Information Filtering: Overview of Issues, Research and Systems”, *User Model. User-Adapt. Interact.*, vol. 11, pp. 203–259, aug. 2001, doi: 10.1023/A:1011196000674.

[8] „7 Advantages of PPC Advertising for Contractors”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.altavistasp.com/2023/01/27/7-advantages-of-pay-per-click-advertising-for-contractors/

[9] „Algorithms :: Model-based Algorithms”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/modelbased.html

[10] J. Leban, „Essentials of recommendation engines: content-based and collaborative filtering”, Medium. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://towardsdatascience.com/essentials-of-recommendation-engines-content-based-and-collaborative-filtering-31521c964922

[11] P. Daru, „Spotify’s Algorithm”, djinit.ai. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://djinit-ai.github.io/2020/04/16/Spotify's-algorithm.html

[12] muffaddal qutbuddin, „An Exhaustive List of Methods to Evaluate Recommender Systems”, Medium. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://towardsdatascience.com/an-exhaustive-list-of-methods-to-evaluate-recommender-systems-a70c05e121de

[13] Z. Deutschman, „Recommender Systems: Machine Learning Metrics and Business Metrics”, neptune.ai. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://neptune.ai/blog/recommender-systems-metrics

[14] Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim, și R. Kashef, „Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities”, *Applied Sciences*, vol. 10, nr. 21, Art. nr. 21, ian. 2020, doi: 10.3390/app10217748.

[15] J. Daikawa, „Building (and Evaluating) a Recommender System for Implicit Feedback”, Medium. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://medium.com/@judaikawa/building-and-evaluating-a-recommender-system-for-implicit-feedback-59495d2077d4

[16] C. Belhekar, „Recommender System Metrics — Clearly Explained”, Medium. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://chaitanyabelhekar.medium.com/recommender-system-metrics-clearly-explained-1f2ba6690216

[17] A. B, „Recommender Systems — It’s Not All About the Accuracy”, Medium. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://gab41.lab41.org/recommender-systems-its-not-all-about-the-accuracy-562c7dceeaff

[18] S. Li, „Building A Recommender System With Implicit Feedback Datasets Using Alternating Least Squares”, Medium. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://actsusanli.medium.com/building-a-recommender-system-with-implicit-feedback-datasets-using-alternating-least-squares-64d4f5ba3c57

[19] „Linear Regression in Machine learning - Javatpoint”, www.javatpoint.com. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.javatpoint.com/linear-regression-in-machine-learning

[20] „Fundamental of Matrix Factorization For Recommender System”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.linkedin.com/pulse/fundamental-matrix-factorization-recommender-system-saurav-kumar

[21] A. Narapareddy, „Recommender system using Bayesian personalized ranking”, Medium. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://towardsdatascience.com/recommender-system-using-bayesian-personalized-ranking-d30e98bba0b9

[22] „Alternating Least Squares | SAP Help Portal”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://help.sap.com/docs/SAP_HANA_PLATFORM/2cfbc5cf2bc14f028cfbe2a2bba60a50/7129de6bddcc490698bee0c2c95e9c73.html

[23] „Welcome to Python.org”, Python.org. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.python.org/doc/

[24] „Project Jupyter Documentation — Jupyter Documentation 4.1.1 alpha documentation”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://docs.jupyter.org/en/latest/

[25] „Loss functions for regression analyses | Machine Learning in the Elastic Stack [8.11] | Elastic”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.elastic.co/guide/en/machine-learning/current/dfa-regression-lossfunction.html

[26] „Retailrocket recommender system dataset”. Data accesării: 7 ianuarie 2024. [Online]. Disponibil la: https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset