Плакат 1

ДОДАТОК А
ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Тема МАР	Дослідження методів вирішення проблеми холодного старту в проектах побудови рекомендаційних систем.
Актуальність	Удосконалений метод забезпечують можливості для підвищення ефективності формування рекомендацій в умовах холодного старту
Об'єкт дослідження	Процеси формування рекомендацій в умовах холодного старту
Мета досліджень	Дослідження методів вирішення проблеми холодного старту в проектах побудови рекомендаційних систем
Задачі досліджень	– аналіз сучасних методів і підходів формування рекомендацій в умовах холодного старту; – дослідження методів побудови рекомендаційної системи в умовах холодного старту; – дослідження особливостей реалізації усунення проблеми холодного старту під час побудови рекомендаційної системи; – опис впровадження рекомендаційної системи до ІТ-продукту.
Методи досліджень	– метод дослідження існуючих методів вирішення проблеми холодного старту, їх формальних описів; – метод дослідження особливостей реалізації усунення проблеми холодного старту під час побудови рекомендаційної системи; – метод дослідження архітектури рекомендаційних систем.
Нові наукові результати	– удосконалений метод побудови рекомендацій в умовах холодного старту з використанням темпоральних обмежень за рахунок введення фільтрації даних на підставі додаткової інформації про користувача
Практична значимість роботи	Проведена апробація запропонованого методу для підвищення ефективності формування рекомендацій в умовах холодного старту

Плакат 2

Архітектура рекомендаційної системи

Архітектура рекомендаційної системи включає в себе наступні складові:
Hadoop-кластер, файлова система HDFS і механізм Hive, інструменти RabbitMQ і Flume, інструмент Sqoop.

Плакат 3

Холодний старт

Проблемою холодного старту називається задача видачі підходящих рекомендацій для нових користувачів. Цю проблему можна розділити на холодний старт для користувачів (нові товари, нові користувачі) і холодний страт для сайтів (рекомендація нових сайтів).

Холодний старт для користувачів обчислюють на основі демографічних даних, які користувачі вказують під час реєстрації на сайтах. Зазвичай це мінімальний набір типу: стать, дата народження та місце розташування.

Плакат 4

Методи побудови рекомендаційних систем в умовах холодного старту

В рекомендаційних системах використовуються наступні методи формування рекомендацій:

1. Метод фільтрації на основі змісту (англ. Content-BasedFiltering);
2. Демографічний метод (англ. Demographic);
3. Метод на основі знань (англ. Knowledge-based);
4. Метод «контекстного багаторукого бандита» (Tangetal)
5. Алгоритм для вирішення проблеми UCoCoS (Sunetal).;
6. Метод на основі тем (TavakolandBrefeld)
7. Метод коллаборативної фільтрації (англ. Collaborativefiltering).
8. **Метод темпоральних обмежень за вхідними даними.**

Плакат 5

Постановка задачі

Для вирішення проблеми холодного старту у проектах побудови рекомендаційних систем необхідно вирішити задачу формування релевантних прогнозів в умовах відсутності вихідних даних.

В ситуаціях «холодного старту» (появи нових користувачів (з порожньою історією переваг) або нових сайтів (не обраних жодним користувачем)). існує практична потреба у побудові рекомендацій для нових користувачів з урахуванням додаткових даних.

Для вирішення проблеми пропонується:

– удосконалити метод побудови рекомендацій в умовах холодного старту з використанням темпоральних обмежень за вхідними даними для його подальшого використання у поставленій задачі.

Даний метод видає в якості рекомендацій в умовах холодного старту найбільш відвідувані сторінки за вказаний проміжок часу. При цьому не враховується додаткова інформація про користувача, що призводить до необхідності аналізу всього списку продажів за вказаний проміжок часу.

Плакат 6

Метод побудови рекомендацій в умовах холодного старту з використанням темпоральних обмежень за вхідними даними

В якості вхідних даних використовуються: журнал продажів, що містить набір трійок $L = \{(u_k, i_j, \tau_n)\}$; часовий інтервал $[\tau_1, \tau_N]$ для вхідних даних; рівень деталізації часу $\Delta\tau$; порогова кількість повторень правила α. Останній параметр використовується при виборі темпоральних обмежень.

Метод включає в себе наступні етапи.

Етап 1. Відбір підмножини записів журналу продажів за заданий інтервал часу $[\tau_1, \tau_N]$.

Етап 2. Узагальнення подій вибору для заданого рівня деталізації часу $\Delta\tau$. На даному етапі значення τ_n узагальнюються відповідно до рівня деталізації - до годин, днів, тощо.

Етап 3. Формування пар (e_j, e_m) послідовного вибору (покупок) користувачів на заданому рівні грануляції часу. Результатом даного етапу є множина R, яка включає пари (e_j, e_m) і кількість повторів цих пар n_{jm} на наборі вхідних даних:

$$R = \{(e_j, e_m, n_{jm}) : e_j X\, e_m, n_{jm} \geq 2\} \quad (1)$$

Етап 4. Формування темпоральних обмежень C для вибору користувачів:

$$C = \{(e_j, e_m) : \forall j \; |\{e_j\}| = |\{e_j, e_m\}| = n_{jm}\} \quad (2)$$

Плакат 7

Етап 5. Відбір темпоральних обмежень для користувачів.

На даному етапі для кожного користувача з кожної множини C **відбирається підмножина обмежень, для яких** $n_{jm} > \alpha$ **:**

$$C_\alpha = \left\{ \left(e_j, e_m \right) : n_{jm} > \alpha \right\} \quad \textbf{(3)}$$

Результатом даного етапу є безліч обмежень по всім користувачам.

Етап 6. Доповнення вхідних даних L записами обмежень згідно (3) для нових, «холодних» користувачів.

$$L^{Cold} = L \bigcup \left\{ (u_k^{Cold}, i_m, \tau_l) : n_{jm} > \alpha \right\} \quad (4)$$

Етап 7. Побудова рекомендацій з використанням традиційних методів, зокрема коллаборативної фільтрації. Результатом даного етапу є перелік рекомендованих товарів для нового користувача, який враховує цикли зміни інтересів відомих користувачів.

Плакат 8

Удосконалений метод побудови рекомендацій в умовах холодного старту з використанням темпоральних обмежень за вхідними даними

Доповнимо етап 5 наступними кроками.

Етап 5.

Крок 1. Аналіз набору даних, які повідомляє про себе користувач в ході реєстрації.

Якщо користувач при реєстрації повідомляє тільки ім'я l і пароль p то необхідно перейти до кроку 3.

Якщо користувач при реєстрації повідомляє про себе необов'язкові додаткові дані (статьs, місце розташування q, вік d), то необхідно перейти до кроку 2.

Крок 2. Виділення підмножини товарів, якими найбільш сильно цікавиться підмножина користувачів з аналогічними додатковими характеристиками, наприклад, всі чоловіки більше 40 років або всі жителі міста Харкова. Результати вибору представити у вигляді тимчасового масиву даних та перейти до кроку 3

Вводимо наступні позначення:

$Event(s,q,d)$- подія, коли користувач U_m вводе додаткові данні s,q,d, де: s - стать користувача;

q – місце розташування користувача;

d – вік користувача, $d_1 < d < d_2$.

$U = \{U_m\}$ – множина користувачів.

$U_m^{lim} = \{s,q,d\}$ - підмножина користувачів яка сформувалася на основі відбору даних по фільтру $Lim^1 = \{s,q,d\}$.

$E = \{E_n\}$ – множина товарів;

Плакат 9

$E_n^{Lim^2} = \{e_j \dots e_n\}$ – підмножини товарів, якими найбільш сильно цікавиться підмножина користувачів $U_m^{Lim^1}$ (на основі фільтру $Lim^2 = \{c, n, z\}$)

$Lim^1 = \{s, q, d\}$ – обмеження за додатковими характеристиками: $\{s, q, d\}$.

$U = \{U_m\}$ – множина користувачів

$U_m^{lim} = \{s, q, d\}$ - підмножина користувачів яка сформувалося на основі відбору даних по фільтру $Lim^1 = \{s, q, d\}$

$E = \{E_n\}$ – множина товарів;

$E_n^{Lim^2} = \{e_j \dots e_n\}$ – підмножини товарів, якими найбільш сильно цікавиться підмножина користувачів $U_m^{Lim^1}$ (на основі фільтру $Lim^2 = \{c, n, z\}$)

$Lim^2 = \{c, n, z\}$ – обмеження за додатковими характеристиками: $\{c, n, z\}$

де c – множина кодів товару;

n– множина назв товару;

z – множина цін.

Правило формування підмножини користувачів:

$$if\ Event(s, q, d)\ then\ Lim^1 = \{s, q, d\}\ then\ U_m^{Lim^1} = \{U_1 \dots U_m\} \quad (1)$$

Правило формування підмножини товарів :

$$if\ U_m^{Lim^1} = \{U_1 \dots U_m\}\ then\ Lim^2 = \{c, n, z\}\ then\ E_n^{Lim^1} = \{e_j \dots e_n\} \quad (2)$$

При цьому виконується відповідність між підмножиною товарів і підмножиною користувачів:

$$U \rightarrow E \quad (3)$$

Крок 3. Відбір темпоральних обмежень для користувачів.

На даному етапі для кожного користувача з кожної множини C відбирається підмножина обмежень, для яких $n_{jm} > \alpha$:

$$C_\alpha = \left\{ \left(e_j, e_m \right) : n_{jm} > \alpha \right\} \quad (5)$$

Результатом даного етапу є безліч обмежень по всім користувачам.

Темпоральні правила формуються не для всіх товарів і користувачів, а тільки для однієї підмножини товарів і однієї підмножини користувачів, які були виділені на кроках 1 і 2.

Плакат 10

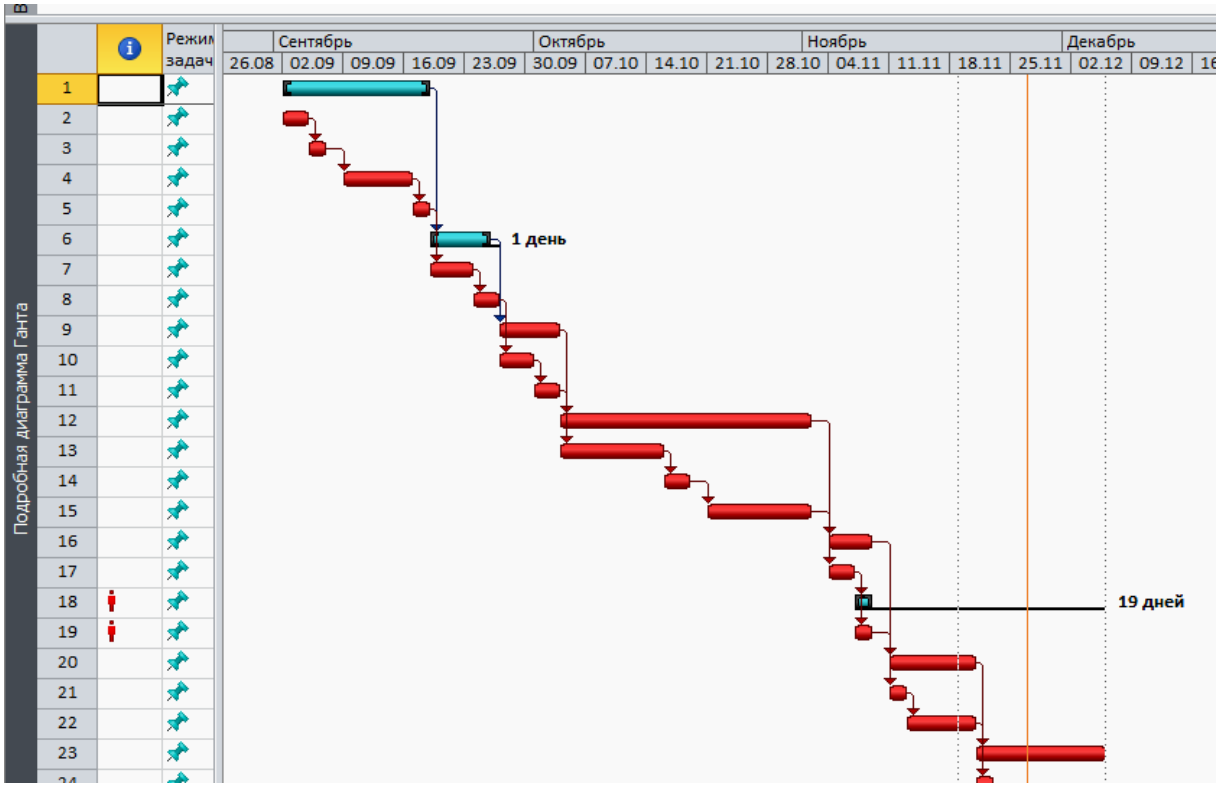
Результати планування IT-проекту

	Название задачи	Длитель	Начало	Окончание	Пред.	Названия ресурсов
1	1. Формування вимог до системи	13 дней	Пн 02.09.19	Ср 18.09.19		
2	1.1 Дослідження предметної області	3 дней	Пн 02.09.19	Ср 04.09.19		Аналітик;Консультант з предметної області
3	1.2 Формування вимог до проекту	2 дней	Чт 05.09.19	Пт 06.09.19	2	Менеджер проекту
4	1.3 Виконання необхідних науково-дослідних робіт	6 дней	Пн 09.09.19	Пн 16.09.19	3	Аналітик
5	1.4 Розробка концепції системи	2 дней	Вт 17.09.19	Ср 18.09.19	4	Аналітик;Консультант з предметної
6	2. Технічне завдання	5 дней	Чт 19.09.19	Ср 25.09.19	1	
7	2.1 Розробка технічного завдання для БД	3 дней	Чт 19.09.19	Пн 23.09.19	5	Технічний письменник
8	2.2 Розробка технічного завдання для створення системи	3 дней	Вт 24.09.19	Чт 26.09.19	7	Технічний письменник
9	3. Проектування БД	5 дней	Пт 27.09.19	Чт 03.10.19	6	
10	3.1 Вибір модалі БД	2 дней	Пт 27.09.19	Пн 30.09.19	8	Розробник БД
11	3.2 Адаптація моделі БД до вимог предметної області	3 дней	Вт 01.10.19	Чт 03.10.19	10	Розробник БД
12	4. Технічний проект	21 дней	Пт 04.10.19	Пт 01.11.19	9	
18	4.1 Розробка БД	8 дней	Пт 04.10.19	Вт 15.10.19	11	Розробник БД
14	4.2 Тестування БД	3 дней	Ср 16.10.19	Пт 18.10.19	13	Розробник БД;Тестувальн
15	4.3 Розробка програмних модулів системи	10 дней	Пн 21.10.19	Пт 01.11.19	14	Програміст
16	5. Тестування	5 дней	Пн 04.11.19	Пт 08.11.19	12	
17	5.1 Тестування модулів системи	3 дней	Пн 04.11.19	Ср 06.11.19	15	Програміст;Тестувальник
18	5.2 Оцінка результатів тестування	2 дней	Чт 07.11.19	Пт 08.11.19	17	Менеджер проекту;Прог
19	5.3 Загальна оцінка проекту	2 дней	Чт 07.11.19	Пт 08.11.19	17	Менеджер проекту
20	6. Оцінка поточного проекту	8 дней	Пн 11.11.19	Ср 20.11.19	16	
21	6.1 Аналіз готовності проекту до експлуатації	2 дней	Пн 11.11.19	Вт 12.11.19	19	Аналітик;Менеджер проекту
22	6.2 Розробка робочої документації для системи	6 дней	Ср 13.11.19	Ср 20.11.19	21	Технічний письменник
23	7. Введення до експлуатації	11 дней	Чт 21.11.19	Чт 05.12.19	20	
24	7.1 Підготовка об'єкта до експлуатації	2 дней	Чт 21.11.19	Пт 22.11.19	22	Менеджер проекту;Прог
25	7.2 Інсталяція продукту	2 дней	Пн 25.11.19	Вт 26.11.19	24	Програміст;Розробник БД
26	7.3 Перевірка працездатності після інсталяції	1 день	Ср 27.11.19	Ср 27.11.19	25	Програміст;Розробник БД
27	7.4 Наповнення БД	2 дней	Чт 28.11.19	Пт 29.11.19	26	Розробник БД
28	7.5 Синхронізація	1 день	Пн 02.12.19	Пн 02.12.19		Програміст;Розробник БД
29	7.6 Тестування	2 дней	Вт 03.12.19	Ср 04.12.19		Тестувальник
30	7.7 Впровадження	1 день	Чт 05.12.19	Чт 05.12.19		Програміст;Розробник БД
31	- підпис погоджувальних документів	1 день	Пт 06.12.19	Пт 06.12.19		Консультант з предметн
32	-усунення недоліків	3 дней	Пт 06.12.19	Вт 10.12.19		Програміст;Розробник БД

Декомпозиція етапів робіт

Плакат 11

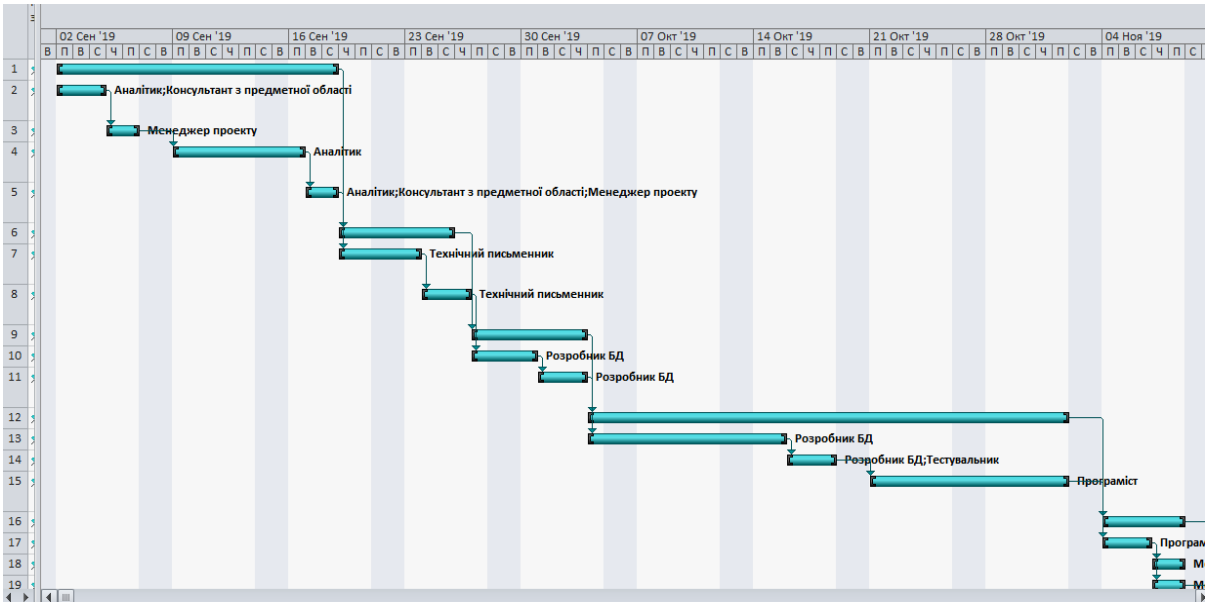
Етапи робіт і ресурси проекту у вигляді діаграми Ганта



Фрагмент докладної діаграми Ганта

Плакат 12

Після формування критичного шляху проекту був виконаний аналіз, результатом якого стала побудова наступної діаграми Ганта:



Плакат 13

Практична апробація методу та тестування результатів

У якості вхідних даних виступає матриця оцінок (табл. 3.1), виставлених користувачами товарів, для зручності товарам привласнені номери 1-9:

Таблиця 3.1 – Матриця оцінок

	1	2	3	4	5	6	7	8	9
U1	5.000	3.000			4.000				
U2	4.000					1.000		2.000	3.000
U3		5.000	5.000						
U4			4.000	3.000		2.000	1.000		

Плакат 14

Косинусна міра для двох векторів:

$$\cos(\vec{x} \cdot \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|_2 x \|\vec{y}\|_2} \quad (3.1)$$

```
def distCosine (vecA, vecB):
    def dotProduct (vecA, vecB):
        d = 0.0
        for dim in vecA:
            if dim in vecB:
                d += vecA[dim]*vecB[dim]
        return d
    return dotProduct (vecA,vecB) / math.sqrt(dotProduct(vecA,vecA)) / math.sqrt(dotProduct(
vecB,vecB))
```

Фрагмент коду обчислення косинусної міри

Плакат 15

```
import math
def makeRecommendation (userID, userRates, nBestUsers, nBestProducts):
    matches = [(u, distCosine(userRates[userID], userRates[u])) for u in userRates if u <> u
serID]
    bestMatches = sorted(matches, key=lambda(x,y):(y,x), reverse=True)[:nBestUsers]
    print "Most correlated with '%s' users:" % userID
    for line in bestMatches:
        print "  UserID: %6s  Coeff: %6.4f" % (line[0], line[1])
    sim = dict()
    sim_all = sum([x[1] for x in bestMatches])
    bestMatches = dict([x for x in bestMatches if x[1] > 0.0])
    for relatedUser in bestMatches:
        for product in userRates[relatedUser]:
            if not product in userRates[userID]:
                if not product in sim:
                    sim[product] = 0.0
                sim[product] += userRates[relatedUser][product] * bestMatches[relatedUser]
    for product in sim:
        sim[product] /= sim_all
    bestProducts = sorted(sim.iteritems(), key=lambda(x,y):(y,x), reverse=True)[:nBestProduc
ts]
    print "Most correlated products:"
    for prodInfo in bestProducts:
        print "  ProductID: %6s  CorrelationCoeff: %6.4f" % (prodInfo[0], prodInfo[1])
    return [(x[0], x[1]) for x in bestProducts]
```

Фрагмент коду

Результат перевірки працездатності коду, отриманий під час тестування, виглядає наступним чином.

```
Most correlated with 'ivan' users:
  UserID:   alex  Coeff: 0.5164
  UserID:  david  Coeff: 0.0667
  UserID:    bob  Coeff: 0.0000
Most correlated products:
  ProductID:      5  CorrelationCoeff: 3.5426
  ProductID:      2  CorrelationCoeff: 2.6570
  ProductID:      3  CorrelationCoeff: 0.4574
  ProductID:      4  CorrelationCoeff: 0.3430
  ProductID:      7  CorrelationCoeff: 0.1143
```

Фрагмент результатів перевірки

Плакат 16

Експертна перевірка методу

А/Б-тест

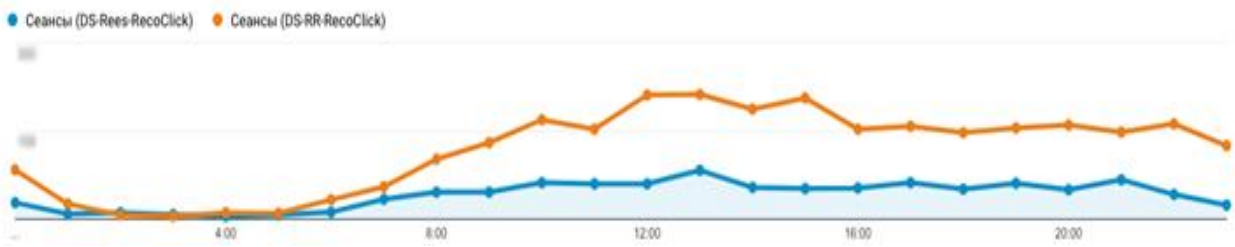
	Конверсія
Рекомендаційна система Інтернет-магазину	20%
Розроблена рекомендаційна система	80%

Результати тесту отримані з Google Analytics – конверсія за сегментами

1. Сегмент А – рекомендаційна система Інтернет-магазину.
2. Сегмент В – розроблена рекомендаційна система.

Таблиця – Результати конверсії.

	03.12.2019	04.12.2019	05.12.2019
Сегмент А	-6,15%	+8,98%	-2,78%
Сегмент В	+3,58%	+10,24%	+2,96%

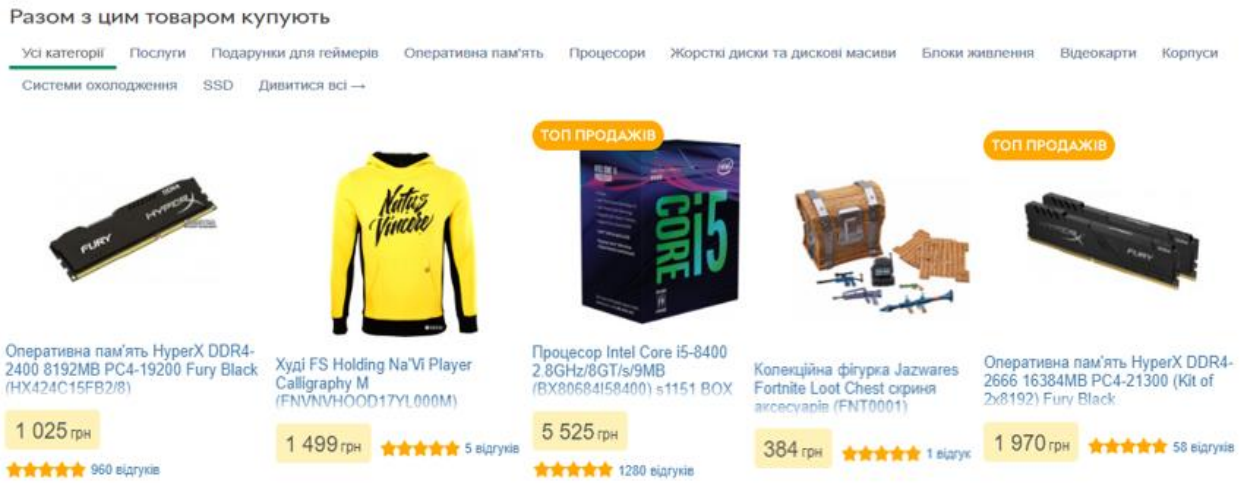


Результати з Google Analytics

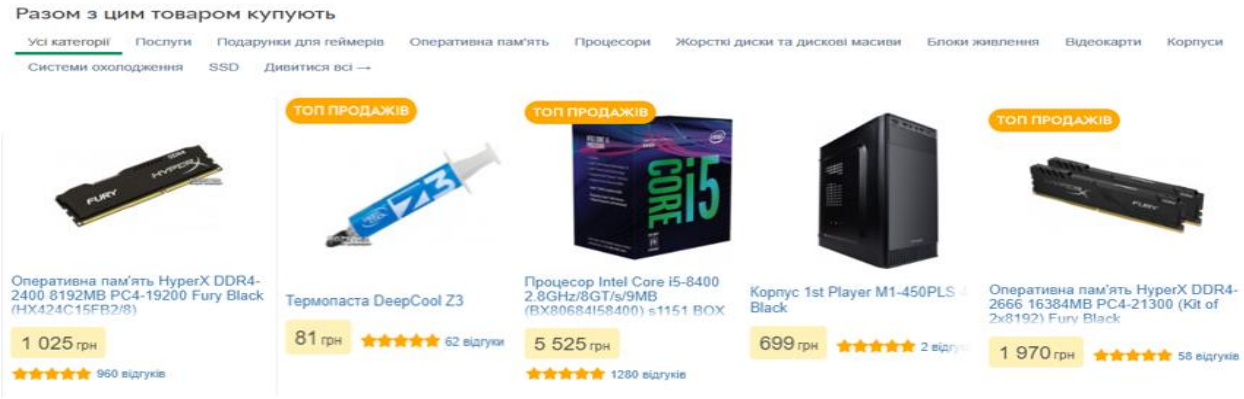
Плакат 17

Приклад без впровадження удосконаленого методу

Рекомендації сформовані обома розглянутими системами для товару «Материнська плата Tomahawk»



Рекомендації системи Інтернет-магазину



Розроблена рекомендаційна система

Плакат 18

Висновки

В результаті виконання магістерської роботи досліджено процес формування рекомендацій в умовах холодного старту.

Проаналізовано існуючі методи формування рекомендацій в умовах холодного старту. Переваги показав методу побудови рекомендацій в умовах холодного старту з використанням темпоральних обмежень за вхідними даними .

Удосконалено метод побудови рекомендацій в умовах холодного старту з використанням темпоральних обмежень за рахунок введення фільтрації даних на підставі додаткової інформації про користувача.

Виконано опис результатів планування ІТ-проекту. Приклади етапів планування наведені за допомогою діаграми Ганта.

Виконано експериментальну перевірку удосконаленого методу. Метод дозволяє покращити показник конверсії від 20% до 80 % в ситуаціях холодного старту.

Плакат 19

Публікації

1. Орехова И.В. Исследование методов решения проблемы холодного старта проектах построения рекомендационных систем [Текст]/ И.В. Орехова // 23-й Міжнародний молодіжний форум «Радіоелектроніка та модуль у XXI столітті». Зб. Матеріалів форуму. Т.6. – Харків: ХНУРЕ. 2019 – 82с. – 83с.

2. Статтю «Удосконалення методу формування рекомендацій за рахунок фільтрації даних на підставі додаткової інформації про користувача в умовах холодного старту» надано до друку в журнал «Біоніка інтелекту»