НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Кафедра автоматизації проектування енергетичних процесів і систем

Лабораторна робота № 6

З дисципліни: «Оброблення надвеликих масивів даних»

**Виконав:**

студент 5 курсу

групи ТР-02мп, ТЕФ

Круглий Д.В.

**Перевірила:**

проф. Федорова Н. В.

Київ – 2020

**Тема:** робота з рекомендаційними системами.

**Мета:** ознайомитися з основними підходами, що покладено в алгоритми рекомендаційних систем. Дослідити вплив поведінки користувача на надані рекомендації. Виконати лабораторне завдання та надати відповіді на запитання.

**Теоретичні відомості**

***Рекомендаційні системи*** - програми, які намагаються передбачити, які об'єкти (фільми, музика, книги, новини, веб-сайти) будуть цікаві користувачеві, маючи певну інформацію про його профілі.

***Типи рекомендаційних систем:***

* Рекомендації по фільмах/книгах/новинах. Пропозиція нового контенту, що підвищує зацікавленість користувачів. Мета полягає в знайомстві користувачів з новим контентом, який може зацікавити.
* Рекомендації по акціях. Пропозиція акцій, які є привабливими та вигідними для клієнтів.
* Рекомендації по продукту. Пропозиція поєднання звичних для клієнта продуктів та новинок.

***Рекомендаційні системи*** порівнюють однотипні дані від різних людей, формують граф інтересів і обчислюють список рекомендацій для конкретного користувача.

***Приклади явного збору даних:***

* запит у користувача оцінки об'єкта за диференційованою шкалою;
* запит у користувача ранжування групи об'єктів від найкращого до найгіршого;
* пред'явлення користувачеві двох об'єктів з питанням про те, який з них краще;
* пропозиція створити список об'єктів, улюблених користувачем.

***Приклади неявного збору даних:***

* спостереження за тим, що оглядає користувач в інтернет-магазинах або базах даних іншого типу;
* ведення записів про поведінку користувача онлайн;
* відстеження вмісту комп'ютера користувача.

**Хід роботи**

1. Мною було досліджено формування стрічки новин соціальної мережі **Facebook**.

Головним завданням новинної стрічки Facebook є демонстрація користувачам «правильного», «особливого» контенту. «Особливий» контент для кожного конкретного користувача може бути різним. З огляду на той факт, що фоловлять користувачі з тими, з ким спілкуються і переписуються, а також на якість контенту і його типі, алгоритм ранжирування Facebook встановлює, в чиїй саме новинній стрічці і на якій позиції виявиться конкретний пост. Тут з'являється цікаве питання: яким чином Facebook встановлює що саме важливо?

Існує ***3 головних правила*** роботи стрічки в Facebook, грунтуючись на яких, соцмережа визначає «особливий» контент для певного користувача:

* Відносини з автором публікації. Тут враховується і час, і кошти взаємодії. Береться до уваги період взаємодії користувача з автором поста за допомогою різних інструментів, у тому числі і Messenger. Все це цілком може змінюватися через якийсь час, тобто поява постів певної людини в вашій стрічці новин буде змінним явищем.
* Тип контенту. Що подобається користувачеві більше - відео, фото або текст? Частіше за інших буде відображатися улюблений тип контенту.
* Якість публікацій. Яка активність була у інших користувачів (позначки «подобається», репости, коментарі)? Своєчасність реакції користувачів також важлива: публікація, до якої був виявлений активний інтерес вчора, на наступний день вже може не бути на першій позиції в новинній стрічці.

Кожен пост отримує оцінку актуальності, яка заснована на трьох розглянутих вище принципах. Кращі з них, які задовольняють цим трьом критеріям, потрапляють в топ. Після показу контенту користувачеві, рейтинг даних постів зберігається. У знову розміщених публікацій також формується свій рейтинг. І, коли у користувача перед очима знову виявиться новинна стрічка, то нові публікації будуть з'являтися над старими відповідно до встановленого рейтингом. З огляду на рейтинг публікацій, Facebook ранжує контент і формує персональну стрічку. Алгоритмічна стрічка враховує десятки чинників, які дозволяють ранжувати якісний контент вище решти. Які фактори впливають на роботу стрічки новин?

***Фактори ранжирування*** стрічки в Facebook:

* перетину інтересів між користувачами (в тому числі відповідно до їх публікацій і лайків);
* регіон, звідки відправлені пости;
* регулярність публікацій у вашій стрічці;
* періодичність публікацій та залученість ваших фанів;
* типи контенту (текст, графіка, відео);
* реакція аудиторії в перші 15-20 хвилин з моменту публікації;
* пріоритет для стрічки віддається відео, які завантажені в соц. мережу;
* глибина перегляду відео;
* опис і тематика відео;

Варто окремо відзначити, що новини, які зменшують охоплення, мають схожі ознаки. У таких новинах використані: хештеги, зовнішні посилання, неунікальні картинки, зображення без тексту, репости без опису змісту поста (1-3 речення).

1. Також я дослідив рекомендаційну систему сервісу **eBay**.

Скласти відповідні додаткові рекомендації щодо елементів, які сприяють перетворенню на eBay, є непростим завданням. eBay - це ринок електронної комерції з 1,2 мільярдами товарів та 179 мільйонами покупців, де користувачі можуть купувати та продавати практично все, що завгодно. Окрім великого масштабу, для цих елементів доступні обмежені структуровані атрибути даних, такі як ISBN, що ускладнює використання традиційних підходів до спільної фільтрації для формування рекомендацій. Інвентар також нестійкий; деякі товари на eBay перелічені лише тиждень і більше ніколи не з’являються. З огляду на всі ці обмеження, важко навіть генерувати подібні рекомендації щодо елемента з урахуванням вихідного елемента (наприклад: seed = iPhone 7 32GB, рекомендація = iPhone 7 64GB). Створення предметів, які доповнюють насіннєвий матеріал, щоб насіння та рекомендовані предмети можна було придбати разом у комплекті, є ще більш складним завданням (приклад: seed = iPhone 7 32GB, рекомендація = iPhone 7). Тут ми описуємо алгоритм додаткових елементів, який ми розробили для вирішення цього завдання.

Алгоритм використовується в модулях на декількох сторінках на сайті eBay, особливо на сторінці товару в смузі мерчандайзингу «Часто купуються разом» під описом насіннєвого товару. Використання *неявних відгуків* користувачів (покупки товарів) лише за допомогою традиційних методів спільної фільтрації не працює на eBay через надзвичайну розрідженість матриці елементів користувача. Якщо говорити неспеціалістами, через дуже велику кількість часто нетривалих предметів наявної інформації про товари, придбані разом одним користувачем, часто недостатньо, щоб з впевненістю давати рекомендації. Однак проведення спільної фільтрації на основі елементів за агрегаціями неявних даних користувачів на рівні елемента має сенс.

Як ви обираєте відповідний рівень агрегування предметів? Природним вибором буде ***агрегування предметів*** на рівні категорії. eBay має таксономію категорій, і всі елементи належать до певної категорії листів у цьому дереві категорій. Ми можемо агрегувати покупки користувачів, щоб сформувати матрицю категорії користувачів, де стовпці представляють категорії листків, а записи в матриці - 1, якщо користувач купував із цієї категорії, або 0 в іншому випадку. Потім ми використовуємо подібність косинусів з відповідними пороговими значеннями, щоб знайти найближчі категорії до найвищих K до вхідної категорії насіння. Пошук найближчих категорій (суміжних категорій) значно обмежує наш простір пошуку можливих позицій кандидатів-рекомендацій, а також зменшує можливість нерелевантних рекомендацій. Усі елементи, які ми рекомендуємо, у всіх наборах відкликання, описаних нижче, походитимуть із цих суміжних категорій.

**Відповіді на контрольні питання**

1. **Назвати основні причини, за якими сервіси застосовують системи рекомендацій.**

Рекомендаційні системи за останнє десятиліття де-факто перетворилися на обов'язковий атрибут OTT-Тб, свого роду «гігієнічний мінімум» в підвищенні лояльності глядачів. Крім утримання існуючих абонентів, рекомендації по вибору контенту залучають нових користувачів і збільшують ARPU, підвищуючи час перегляду і платежі від продажу контенту.

1. **Назвати суттєві. характеристики рекомендаційних систем.**

Дві основні стратегії створення рекомендаційних систем - *фільтрація на основі вмісту* і *колаборативна фільтрація*. При фільтрації на основі змісту створюються профілі користувачів і об'єктів, профілі користувачів можуть включати демографічну інформацію або відповіді на певний набір питань, профілі об'єктів можуть включати назви жанрів, імена акторів, імена виконавців та іншу атрибутивну інформацію в залежності від типу об'єкта. Наприклад, в Music Genome Project музичний аналітик оцінює кожну композицію по сотням різних музичних характеристик, які можуть використовуватися для виявлення музичних уподобань користувача. При колаборативної фільтрації використовується інформація про поведінку користувачів в минулому - наприклад, інформація про покупках або оцінках. В цьому випадку не має значення, з якими типами об'єктів ведеться робота, але при цьому можуть враховуватися неявні характеристики, які складно було б врахувати при створенні профілю. Основна проблема цього типу рекомендаційних систем - «холодний старт»: відсутність даних про недавно з'явилися в системі користувачів або об'єктах.

1. **Навести приклади явного і неявного збору даних про користувача.**

Приклади *явного збору* даних:

- запит у користувача оцінки об'єкта за диференційованою шкалою;

- запит у користувача ранжування групи об'єктів від найкращого до найгіршого;

- пред'явлення користувачеві двох об'єктів з питанням про те, який з них краще;

- пропозиція створити список об'єктів, улюблених користувачем.

Приклади *неявного збору* даних:

- спостереження за тим, що оглядає користувач в інтернет-магазинах або базах даних іншого типу;

- ведення записів про поведінку користувача онлайн;

- відстеження вмісту комп'ютера користувача.

1. **Навести приклади популярних ресурсів, де спостерігається вплив рекомендаційної системи на запити користувача.**

Приклади популярних ресурсів, де спостерігається вплив рекомендаційної системи на запити користувача: Amazon, eBay, Netflix, Megogo, Deezer, Last.fm, LinkedIn, YouTube.

1. **На які фактори спирається система рекомендацій YouTube?**

Кожну хвилину на YouTube з'являється мінімум 300 годин нового відеоконтенту різного змісту. Щомісяця на платформу заходить близько двох мільярдів користувачів, які переглядають понад мільярд годин відео. Для обробки величезного обсягу інформації компанія активно використовує можливості штучного інтелекту, щоб спростити деякі операції. YouTube з щохвилинної завантаженням великої кількості відео відрізняється однією з найбільш просунутих систем рекомендацій, створеної на основі штучного інтелекту. Вона набагато відрізняється від механізмів, які використовуються на Netflix, Hulu, або Spotify, забезпечуючи обробку постійно оновлюваного контенту і формування рекомендацій в режимі реального часу.

Розроблена система складається з двох частин. Перша є нейронною мережею, яка призначена для генерації кандидатів на основі історії переглядів користувачів YouTube. Це дозволяє забезпечити широку персоналізацію з подальшою фільтрацією контенту за такими ідентифікаторами, як кількість і час переглядів відео, демографічна інформація та пошукові запити. Друга частина - мережа ранжирування, яка привласнює оцінку для кожного відео, використовуючи «широкий набір функцій, що описують користувачів і ролик». Двохрівневий підхід дозволяє системі обробляти мільйони відео, рекомендуючи змістовний корисний контент. Головний показник корисності рекомендацій - час перегляду. Система дуже ефективна для платформи і рекламодавців, однак, може не влаштовувати користувачів, оскільки не завжди тривалість перегляду є підтвердженням зацікавленості.

1. **Назвати цінності, які покладає на своїх користувачів Facebook.**

У Facebook є офіційний документ у відкритому доступі. Він називається Facebook News Feed Values - цінності новинної стрічки. Це корпоративна книга в компанії - документ, який описує, що покладено в основу Facebook, і чого він очікує від користувачів.

«***Ви повинні інформувати***». Люди очікують, що історії в їх стрічці будуть давати їм щось корисне. Правда, люди різні і те, що одна людина знаходить інформативним або цікавим, не буде таким для іншої. Facebook постійно працює над персоналізацією контенту.

«***Ви повинні розважати***». Facebook дає різні види розваги: перегляд прямих ефірів, обмін фотографіями з друзями або можливість стежити за життям публічних осіб. Відгук аудиторії на пости добре відслідковується через емотікони (англ. emoticon, emotion icon — іконка з емоцією). Емотікони створені для того, щоб відстежувати алгоритм і дати зрозуміти, який контент викликає в аудиторії різні емоції. Тому важливо, щоб контент був різноманітним: пізнавальним, конкретним, емоційним.

«***Платформа для всіх ідей***». Класична цінність для всіх американських сервісів і бізнесів - контент не ображає жодні права, не обмежує меншини. В стрічці може з'являтися все, що не заборонено правилами.

«***Автентичні комунікації***». Чим конкретніше, чесно і не упереджено користувач висвітлює свою діяльність, тим вище буде пріоритет від алгоритму. Фейкові новини і вражаючі заголовки на кшталт «Як схуднути за допомогою соди», «Неймовірна новина - рибалки витягли з озера ЦЕ» будуть знижуватися в рейтингу пошуку.

«***Ви самі контролюєте свій досвід***». Кожен користувач може регулювати пости, що з‘являються в його стрічці.

«***Постійні оновлення***». Команда Facebook постійно працює над покращеннями алгоритмів розумної стрічки. Анонси новинок або важливих змін в алгоритмах висвітлюються через посилання FYI (For Your Attention), Facebook робить все, щоб перебувати в стрічці було максимально комфортно.

1. **Які фактори суттєво впливають на формування стрічки новин Facebook для конкретного користувача?**

Найбільше на стрічку новин Facebook впливають запити, які користувач вводить у пошук соціальної системи. Тобто, у випадку пошуку якоїсь інформації: статей на тему «карантин в Україні», варто очікувати найближчим часом (декілька днів) suggested публікації на цю ж саму тему. Не варто забувати також про систему лайків та дизлайків в Facebook (та у інших системах). Очевидно, що канали, статті, пости, на які користувач підписаний або на яких користувач ставить лайки (з деякою періодичністю) приймають участь у конкурентній боротьбі за recommended stuff.

1. **Назвати формати взаємодії користувача з оточенням та їх вагомість для ранжирування.**

Зараз алгоритм показує в новинній стрічці пости таким чином, щоб кожен користувач бачив їх в тому порядку, в якому вони йому максимально сподобаються і будуть цікаві. І для цього система використовує певні чинники ранжирування.

Фактори ранжирування - це дані про минулу поведінку певного користувача і поведінку всіх користувачів платформи. Алгоритм зважує багато чинників перед тим, як вирішити, чи показувати користувачеві певний пост.

Важливі категорії факторів ранжирування:

* З ким зазвичай взаємодіє користувач.
* Тип медіа в пості (відео, посилання, фотографія тощо).
* Популярність поста.

1. **За яким принципом працює алгоритм Facebook для рекомендації друзів?**

Для списку “Люди: яких ви можете знати”, я знайшов два підбору рекомендованих осіб:

- Очевидно за спільними друзями. Вірогідність того, що ти колись бачив, спілкувався друга твого друга набагато вища, ніж із випадковою людиною.

- У цій стрічці потрапляють особи, з якими я не маю ні одного друга у спільних. Але я помітив, що Facebook використовує географічне положення для формування списку. Тобто, люди, які були одночасно в одному (майже в одному) і тому самому місці, можливо могли завести бесіду.

1. **Які фактори суттєво впливають на формування стрічки новин Instagram для конкретного користувача?**

Щоб зробити свої рекомендації, система Explore починає з перегляду «насінних облікових записів», що є обліковими записами, з якими користувачі взаємодіяли раніше, сподобавшись або зберігши їхній вміст. Він ідентифікує подібні до них рахунки, і з них відбирає 500 частин вмісту. Ці кандидати відфільтровуються для видалення спаму, дезінформації та "ймовірного вмісту, що порушує політику", а решта публікацій класифікується на основі імовірності взаємодії користувача з кожною з них. Нарешті, 25 найкращих публікацій надсилаються на першу сторінку вкладки "Огляд" користувача.

Тут слід зазначити кілька речей. По-перше, Instagram не є повністю прозорим щодо свого процесу. Немає детальної інформації про те, які сигнали використовуються для ідентифікації спаму чи дезінформації, і це не надто дивно, враховуючи, що пояснення цього допоможе людям, які хочуть поширювати такий вміст. Компанії також незрозуміло, якою мірою машинне навчання використовується для фільтрації невідповідного вмісту - ключова деталь, враховуючи, що Facebook часто представляє ШІ як чарівну кульку для модерування (тоді як експерти не погоджуються).

**Висновок:** під час виконання лабораторної роботи я ознайомився з основними підходами, що покладено в алгоритми рекомендаційних систем. Дослідив вплив поведінки користувача на надані рекомендації.