Слајд 1

У данашњем свету где свака врста пословања налази неку примену и везу са интернетом, електронска трговина је постала један од најкоришћенијих услуга које нуди интернет. Уместо да корисници морају да иду лично у продавнице, могу из топлине својих домова да пазаре неопходне производе. Е-commerce апликације пружају лак приступ широком спектру производа и услуга. Баш због олакшавања процеса куповине производа све више корисника користе могућност електронске трговине чиме се број корисника и трансакција драстично повећао. Праћењем понашања корисника, генеришу се велике количине искористивих информација у виду претрага, куповина и оцењивања. Велики изазов је постао да се нађе начин да се те информације искористе како би се побољшало корисничко искуство тиме што би се нудиле персонализоване препоруке производа самим корисницима.

Доказано је да уколико се корисницима нуде препоруке на основу њихових интересовања повећава се вероватноћа куповине и повратка корисника на исти сајт, при чему су и корисници много задовољнији. Међутим, имплементација овакве врсте система није лака и доноси велики број изазова, нарочито када је број података који треба обрадити велики. Систем мора да прати корисничко понашање константно како би сакупио што више информација о њему како би повећао прецизност и тачност препорука.

Слајд 2

Системи за давање препорука су информационо-филтрирајући алати који помажу корисницима да пронађу релевантне производе или информације у великим базама података. Главна функција ових система је генерисање садржаја према корисничким преференцијама. Пример таквог система нуди Аmazon који аутоматски препоручује производе кориснику базиране на предходним куповинама, оценама и интеракцијом корисника са производима. Такође се систем за давање препорука може посматрати и као страшки алат за доношење одлука у системима који обрађују велику количину података.

Порастом корисника и употребе интернета крајем деведесетих година креће и развој првих система за давање препорука. Као и сваки други пројекат који је у развоју, тако и системи за давање препорука нису били савршени, били су јако прости и препоручивали су корисницима најпродаваније производе. Временом су схватили да таква врста препорука не доприноси много повећању продаје и да би систем био много кориснији да се кориснику препоручују њему занимљиви производи, а не глобално занимљиви производи

Слајд 3

Имплементација система за препоруке у е-commerce апликацијама нуди бројне предности које директно утичу на раст продаје, унапређење корисничког искуства и оптимизацију пословања. Ове предности укључују повећање прихода, лојалности корисника и ефикасније трошење маркетиншких ресурса.

1. Повећање продаје

Kao што је већ наглашено системи за давање препорука се креирају да би се повећала количина продаје производа. Анализе и истраживања тржишта креирали су податак да коришћење квалитетних система за давање препорука повећава број купаца за чак 22,66% [5]. Такође постојећи купци могу докупити додатни производ ако систем препоручи производ за који предпоставља да ће се свидети кориснику. [4]

2. Побољшање корисничког искуства и лојалности

Системи за давање препорука доприносе побољшању корисничког искуства тиме што скраћују време претраге и олакшавају корисницима проналажење жељених производа. Препоруке прилагођене преференцијама корисника не само да повећавају задовољство, већ и подстичу лојалност, јер се корисници враћају на платформе које боље разумеју њихове потребе. Корисник ће пре изабрати платформу на којој ће брже и лакше задовољити своје жеље. [6]

3. Унапређење пословања и смањење трошкова компаније

Обзиром да се обрађују подаци о корисницима и њиховим преференцијама, компаније имају увид у производе које занимају већину корисника. Са тим знањем компаније знају у ком правцу треба да иде њихово пословање како би се продаја повећавала односно који тип производа требају више набављати и продавати а који мање. [5]

Ови разлози јасно указују на важност имплементације система за препоруке у модерним е-commerce апликацијама, где персонализација и ефикасност постају кључни фактори у одржавању конкурентности и задовољства корисника.

Слајд 4

Типови система за давање препорука:

1. Колаборативно филтрирање (енгл. Collaborative filtering)

Колаборативно филтрирање је један од најранијих и најчешће коришћених приступа у системима за препоруке. Модел се заснива на идеји да се на основу корисничких преференција и понашања генеришу препоруке.

2. Филтрирање по садржају (енгл. Content-based Filtering)

Филтрирање по садржају (слика 2) представља технику која анализира историју корисника, концетришући се на производе са којима је корисник имао интеракцију у виду куповина или оцењивања. Проналази сличне производе анализираним производима на основу сличности у њиховим карактеристикама, описима, атрибутима.

3. Хибридне методе (енгл. hybrid methods)

Хибридне методе (слика 3) настале су као комбинација предходних техника. Комбиновањем метода добијају се прецизније препоруке и повећава се задовољство крајњих корисника. Такође хибридним методама смањују се недостаци појединачних метода и комбинују се предности.

Слајд 5

Колаборативно филтрирање

Колаборативно филтрирање је један од најранијих и најчешће коришћених приступа у системима за препоруке. Модел се заснива на идеји да се на основу корисничких преференција и понашања генеришу препоруке. Колаборативно филтрирање (слика 1) се може поделити на два приступа: базиран на меморији (memory-based) и базиран на моделу (model-based). Колаборативно филтрирање базирано на меморији је метод који примењује различите алгоритме над подацима које систем поседује о корисничкој историји, док с друге стране колаборативно филтрирање базирано на моделу је приступ који користи машинско учење за креирање модела који предвиђају преференције корисника. Постоје две врсте memory-based колаборативног филтрирања колаборотивно филтрирање на основу корисника (енгл. user-based filtering) и колаборативно филтрирање на основу производа (енгл. content-based filtering).

Главна идеја колаборативног филтрирања је складиштење и анализа података о оценама и куповинама корисника. Систем користи ове податке да пронађе кориснике са сличним интересовањима или производе сличне онима за које је корисник већ показао интересовање. На основу тих информација, генеришу се препоруке које одговарају специфичним потребама и преференцијама корисника. [3]

Колаборативно филтрирање развијено је почетком деведесетих година и брзо је постало стандард у системима за давање препорука. Као стандард у колаборативном филтрирању устаљиле су се две технике: колаборативно филтрирање засновано на кориснику и колаборативно филтрирање засновано на производу. [4]

Слајд 6

На основу корисника

Ова метода се базира на идеји да се корисничке препоруке генеришу на основу сличности међу корисницима. Главна сврха овог приступа је да идентификује кориснике са сличним интересовањима како би се предвидело шта би тренутни корисник могао да воли на основу преференција њему сличних корисника. [2]

Кориснички профил се користи као основа за препоруке. Када корисник користи систем, његов профил се упоређује са профилима других корисника како би се пронашли „суседи" који су имали сличне преференције у прошлости. На основу оцена које су ти слични корисници дали различитим производима, систем може да процени које производе би тренутни корисник могао да воли, а које још није видео.

Слајд 7

На основу производа

Item-based collaborative filtering препоручује производе на основу сличности између производа. Систем анализира које су производе корисници куповали или оцењивали слично и користи те податке да предложи сличне производе другима. Фокус је на односима међу артиклима, не корисницима.

слајд 8

Проблем хладног старта

Проблем хладног старта јавља се када систем не може да генерише ефикасне препоруке због недостатка података. Овај проблем је нарочито изражен код нових корисника или нових производа, где систем нема довољно информација (корисничких преференција и оцена) како би дао одговарајуће препоруке. Постоји неколико начина да би се пребродили ови проблеми:

Једно од решења је да се од корисника приликом регистрације тражи да унесе одређене преференције или да оцени неке производе како би систем имао довољно података на основу којих може да кориснику понуди одговарајуће препоруке.

Прикупљање демографских података, као што су године, пол или интересовања корисника, може помоћи у давању одговарајућих препорука у почетним фазама коришћења система.

Пробем реткости података

Проблем реткости података јавља се када активни корисници оцене само мали број производа. У великим е-commerce системима, где корисници обично оцењују само мали проценат производа, систем може давати неквалитетне препоруке услед недостатка података. Неке од техника за превазилажење овог проблема укључују :

Демографско филтрирање – коришћење информација о корисницима као што су старост, пол или интересовања, како би се употпунили подаци.

Singular Value Decomposition (SVD) – техника смањења димензионалности која омогућава боље коришћење доступних података у случају малог броја оцена.

Model-based технике – алгоритми који користе моделе, попут кластеровања, за груписање корисника или производа ради генерисања прецизнијих препорука чак и у условима ретких података.

Проблем скалабилности

Са порастом броја корисника и производа, долази до проблема скалабилности. У системима са великим бројem корисника и производа, генерисање препорука у реалном времену постаје тешко због потребе за обрадом велике количине података, па су самим тим алгоритми спори. Два уобичајена приступа за решавање овог проблема су:

Смањење димензионалности – методе попут SVD могу се користити за смањење комплексности података, омогућавајући бржу обраду.

Технике засноване на кластеровању – уместо да се претражује цела база података, корисници се могу груписати у мање кластере како би се брже пронашле препоруке за специфичне групе.

слајд 9

Филтрирање по садржају је техника која се фокусира на карактеристике производа са којима је корисник имао интеракцију. Филтрирање по садржају се базира на идеји да ако је корисник показао интересовање за производ који садржи одређену вредност неког атрибута (на пример ако су у питању филмови, и ако је корисник љубитељ акција) да се кориснику нуде производи који такође садрже исту вредност датог атрибута (слика 22). [2]

За разлику од колаборативног филтрирања, филтрирање по садржају се базира искључиво од корисничког профила и његових преференција, не узимају се у обзир други корисници. Кориснички профил се креира на основу историје активности корисника, тј. креира се на основу производа које је корисник на неки начин означио као фаворите (купио их је, фаворизовaо…). Због изолованости само на личне преференције филтрирање по садржају нема проблема, уколико не постоје информације о активностима других корисника и потпуно правилно функционише и уколико у систему постоји само један корисник. Такође је отпоран и на проблеме хладног старта у односу на нове производе, јер уколико су они слични фаворизованим производима корисника, могу бити препоручени кориснику. [12]

Карактеристике производа могу бити издвојени атрибути, али исто тако могуће да производ садржи само опис. Технике филтрирања по садржају се често примењују на текстуалне препоруке, као што су чланци или wеb странице. Систем анализира кључне речи и друге релевантне карактеристике предмета како би пронашао сличности са предметима које је корисник раније високо оценио. Иако ова метода захтева познавање карактеристика предмета, она пружа флексибилност у персонализацији и прилагођавању препорука према преференцијама корисника које се временом мењају.

За утврђивање сличности између предмета користе се различите функције. Једноставан приступ би био проверавање да ли је производ описан неким атрибутом задовољава корисничке преференције односно да ли се проверавани и фаворизовани производ поклапају по вредности атрибута (да ли су оба филма акције, уколико је реч о филмовима). Сличност би могла бити 0 (не постоји преклапање) или 1 (потпуно преклапање). Међутим, систем може користити и сложеније метрике, као што је Dice коефицијент, који мери сличност на основу преклапања скупа кључних речи између различитих предмета [2]. За два производа описаних скупом кључних речи Dice кеофицијент сличности рачуна се на основу формуле са слике 23.

Међутим у неким случајевима није могуће описати производ скупом атрибута, на пример уколико се креира систем за препоруке књига информације попут аутора и издавача нису део директног садржаја књиге, али су значајни за описивање садржаја. Па се због тога садржај често представља као скуп кључних речи које су аутоматски генерисане из текста или описа. На основу кључних речи документ се може анализирати у облику вектора у вишедимензионалном простору. Сваки документ може бити представљен бинарним вектором, где 1 означава присуство одређене кључне речи, а 0 њено одсуство. Мана овог приступа је игнорисање учесталости појављивања речи и дужину документа. [2]

Да би се превазишли ти проблеми, користи се ТF-IDF метода. ТF (енгл. term frequency) означава колико често се одређена реч појављује у документу, при чему се у обзир узима и дужина документа. IDF (енгл. inverse document frequency) смањује тежину оних речи које се често појављују у свим документима (везника, речца, прилога…). Комбиновањем ова два параметра, ТF-IDF трансформише документ у вектор који описује његов садржај, смањујући утицај уобичајених речи и истичући оне које су специфичне и битне за препоруку. Као резултат, систем за давање препорука може упоредити векторе корисничког профила и векторе производа са којима корисник није имао додир како би генерисао препоруке базиране на сличностима у садржају истих

Проблеми с којима се суочава

1. Плитка анализа садржаја

Системи који користе филтрирање по садржају често се ослањају на текстуални опис производа, што може бити недовољно за прецизно одређивање да ли тај производ треба препоручити или не. Систем не може разликовати квалитетно написане описе од лошијих, јер се ослања само на кључне речи. Поред упитног квалитета описа производа такође могу представљати проблеме и кратки описи или предмети са мало атрибута због тога што отежавају алгоритму да направи одговарајуће корисничке преференције. Такође проблем представљају и нетекстуални подаци јер системи обично не узимају у разматрање мултимедијалне податке, што може бити кључно за препоруке. [2]

1. Превише специјализоване препоруке

Велики проблем филтрирања по садржају је што се нуде производи који су слични производима с којима је корисник већ имао интеракцију. Самим тим препоруке постају монотоне и из препорука се изостављају проиводи који нису слични досадашњим преферираним производима корисника, али који такође могу да буду занимљиви. Технике које се користе како би се у некој мери ублажио наведени проблем су „тематске диверсификације” или убацивања „насумичних” препорука. [2]

1. Хладан старт на нивоу корсисника

Иако хладан старт на нивоу нових производа не представља проблем филтрирању по садржају, хладан старт на нивоу новог корисника негативно утиче на алгоритам. Проблем настаје кад нови корисник нема довољно оцењених производа, па је тешко креирати одговарајући кориснички профил што резултира не тако добрим препорукама. Решење овог проблема је као и код колаборативног филтрирања да се приликом регистровања новог корисника, самом кориснику наметне да оцени неколико производа како би се креирао почетни кориснички профил који ће се временом усавршити

Хибридни системи за давање препорука

Хибридни системи за давање препорука су комбинација различитих техника, најчешће филтрирања по садржају и колаборативног филтрирања. Главна предност хибридног система је његова способност да искористи добре стране обе технике, чиме се постиже побољшање прецизности препорука и обезбеђује већи степен прилагодљивости потребама корисника. На пример, систем може користити филтрирање по садржају за препоруку производа сличних онима које је корисник већ оценио и колаборативно филтрирање за препоруку производа које су слични корисници оценили позитивно и да комбинацијом та два резултата генерише најадекватније препоруке. Поред идеје да се искористе само врлине обе технике, хибридни системи су настали да би се смањили проблеми које доноси свака техника појединачно. Хибридни модели за давање препорука комбинују ове две технике како би надокнадили недостатак података и побољшали перформансе система. Комбинација наведених техника у хибридном моделу може се остварити на неколико начина. На пример, систем може засебно имплементирати алгоритаме и комбиновати резултате, резултат једног алгоритма проследити као улаз другог или обрнуто.

duck db

DuckDB је аналитичка, уграђена, релациона база података која је дизајнирана за рад са великим количинама података. DuckDB је релативно нова база података која је брзо стекла популарност због своје једноставности и ефикасности у аналитичким упитима. Први пут је представљена 2019. године, а њен циљ је да реши проблеме с којима се суочавају традиционалне базе података при обради великих количина података у оквиру аналитичких задатака. DuckDB је често упоређиван са SQLite-ом, јер се обе базе фокусирају на минималну инсталацију и лакоћу коришћења, али за разлику од SQLite-а, DuckDB је оптимизован за онлајн аналитичку обраду података (OLAP). [19][27]

Једна од кључних предности DuckDB-а је то што нема екстерне зависности, што значи да не захтева комплексне инсталације или одржавање сервера. Све се одвија унутар хост процеса, што олакшава интеграцију и побољшава брзину преноса података. DuckDB је посебно дизајниран за обраду великих аналитичких упита, као што су агрегација и спајање табела. Његова архитектура омогућава ефикасно извршавање упита захваљујући колонарно-векторизованом механизму за извршавање упита, што га чини бржим у поређењу са традиционалним базама података као што су PostgreSQL или MySQL. Такође DuckDB је бесплатан и отвореног кода што га поред осталих предности чини јако пожељним алатом. [19][27]

DuckDB омогућава обраду података који су складиштени на различите начине. DuckDB има подршку за учитавање CSV, JSON, Parquet и Apache Arrow датотека, али исто тако могуће је повезати DuckDB са неком другом постојећом базом података над чијим ће се подацима извршавати упити, као што су Postgres или SQLite. Поред обраде DuckDB омогућава и складиштење података у већ поменутим врстама датотека као и у базама података. [18]

## Разлози за коришћење DuckDB базе података

1. Брзина, ефикасност и смањење трошкова

DuckDB је посебно ефикасан када се ради о локалној обради података. DuckDB омогућава локалну обраду података чиме се губи потреба за online cloud сервисима за аналитичку обраду података. Овом карактеристиком DuckDB-а се јако убрзава обрада података и смањују се трошкови јер се не користе cloud сервиси попут AWS Athena или BigQuery-а. Уштеда се највише примeћује током обраде велике количине података. [18]

1. Једноставна инсталација

DuckDB се једноставно инсталира у различитим окружењима. Поред једноставности инсталације, DuckDB не захтева велику количину ресурса. Има подршку за велики број програмских језика као што су: C, C++, Java, Node.js, Python… Такође је доступан и за оперативне системе Windows, Linux, macOS. [18]

1. Напредне SQL функционалности

Самим тим што је DuckDB релациона база података, поседује SQL упитни језик, при чему има и неке додатне проширене функције у односу на стандардни SQL који је познат већини програмера. Проширени SQL омогућава да се сложени упити обављају брже и са мање кода. [18]

1. Анализа приватних података

Самим тим што је DuckDB ефикасан за локалну обраду, значи да се њиме могу ефикасно обрађивати локални приватни подаци са већом сигурношћу јер се не обрађују online. Ово је веома битна предност у системима у којима се обрађују осетљиви и тајни подаци. [18]

1. Дистрибуирана обрада података

Иако је DuckDB првенствено намењен локалној аналитици, сваког дана се DuckDB развија и као алат за дистрибуирану обраду података. Пројекат MotherDuck омогућава коришћење DuckDB-а и у cloud окружењима, што даје додатну флексибилност када је потребно анализирати податке са више различитих извора. [18]

## Стандардне примене DuckDB-a

1. Edge computing

DuckDB је ефикасан за анализу података директно на месту где се они прикупљају, као што су паметни бројачи или IoT сензори. Ова локална анализа омогућава брже добијање резултата и смањује потребу за слањем великих количина података у cloud, што помаже у уштеди трошкова преноса и обраде. [19]

1. Анализа великих скупова података

DuckDB је веома ефикасан за анализу великих скупова података, као што су log датотеке, без потребе за њиховим премештањем у cloud. DuckDB омогућава да се подаци анализирају директно тамо где се налазе, што може значајно смањити трошкове и време обраде. [18]

1. Data science и машинско учење

DuckDB може олакшавати рад и анализу јер омогућава бржу припрему, филтрирање и агрегацију података. Због своје брзине обраде великих количина података погодан је за системе који се базирају на машинском учењу, као што су системи за препоруку садржаја. [18]

1. Дистрибуирана анализа података

DuckDB се може користити за анализу података који су расподељени између cloud-а, edge мреже и локалних уређаја. Пројекат MotherDuck омогућава коришћење DuckDB-а у оба окружења тачније у cloudu и локално што омогућава ефикасну обраду података на различитим нивоима. [18]

1. Трансформација и чување података

DuckDB може бити коришћен за трансформацију и припрему података пре него што се обави анализа. Омогућава ефикасну промену назива колона, типова података и вредности, као и рад са угњежденим документима, претварајући их у релационе структуре које су лакше за анализу. [18]