Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

факультет прикладних інформаційних технологій та електроінженерії
(повна назва факультету)
кафедра автоматизації технологічних процесів і виробництв
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: «Розроблення інформаційної системи для аналізу баз даних та прийняття оптимальних рішень з використання наявних ресурсів металургійних підприємств »

Виконав(ла): студент(ка) VI курсу, групи КАм-61
спеціальності 151 «Автоматизація та комп'ютерно-інтегровані технології»
(шифр і назва спеціальності)

(підпис) Сенко В.Ю. (прізвище та ініціали)

Керівник (підпис) Марущак П.О. (прізвище та ініціали)

Нормоконтроль (підпис) Козбур І.Р. (прізвище та ініціали)

Завідувач кафедри (підпис) Савків В.Б. (прізвище та ініціали)

Рецензент (підпис) Микитишин А.Г. (прізвище та ініціали)

Тернопіль
2022

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет прикладних інформаційних технологій та електроінженерії
(повна назва факультету)

Кафедра автоматизації технологічних процесів і виробництв
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
Савків В.Б.
(підпис) (прізвище та ініціали)
« ___» _____2022р.

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня магістр
(назва освітнього ступеня)

за спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології»
(шифр і назва спеціальності)

студенту Сенко Володими Юрійович
(прізвище, ім'я, по батькові)

1. Тема роботи «Розроблення інформаційної системи для аналізу баз даних та прийняття оптимальних рішень з використання наявних ресурсів металургійних підприємств »

Керівник роботи проф. Марущак П.О.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджена наказом ректора від «11» листопада 2022 року № 4/7-895

2. Термін подання студентом завершеної роботи 20 грудня 2022 року

3. Вихідні дані до роботи База даних виробів підприємства, технологічні вимоги до рекомендаційної системи.

4. Зміст роботи (перелік питань, які потрібно розробити)
1) аналітична частина; 2) технологічна частина;
3) конструкторська частина; 4) науково-дослідницька частина; 5) спеціальна частина
6) Безпека життєдіяльності, основи охорони праці.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
Презентація кваліфікаційної роботи аркушів формату А4

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Спеціальна частина	*проф. Марущак П.О.*		
Охорона праці	*доц. Тотосько О.В.*		
Безпека в надзвичайних ситуаціях	*ст.викл Клепчик В.М.*		

7. Дата видачі завдання *«____» _____20_____ р.*

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи бакалавра	Термін виконання етапів кваліфікаційної роботи	Примітка
1.	*Аналітична частина*	*12.11.2022*	
2.	*Проєктна частина*	*30.11.2022*	
3.	*Спеціальна частина*	*05.12.2022*	
4.	*Безпека життєдіяльності, основи охорони праці*	*10.12.2022*	
5.	*Оформлення графічної частини та пояснювальної*	*15.12.2022*	
	записки		
6.	*Захист кваліфікаційної роботи*	*23.12.2022*	

Студент _____ *Сенко Володимир Юрійович*
(підпис) (прізвище та ініціали)

Керівник кваліфікаційної роботи _____ *Марушак Павло Орестович.*
(підпис) (прізвище та ініціали)

ЗМІСТ

ВСТУП

Ця робота присвячена розробці рекомендаційних систем для металургійних підприємств з великою номенклатурою виробів та приладів обслуговування даного підприємства з використанням кластеризації баз даних графів. Метою даної роботи було дослідити використання графових баз даних як методу розробки системи рекомендацій.

Із зростанням об’ємності виробництва та кількості виробів що виготовляє дана промисловість а також устаткування яке використовується на даному виробництві корисно розробити систему, яка може надавати рекомендації для спрощення пошуку та організації зв'язків між виробами та устаткуваннями, також дана система може використовуватися для управління наявними матеріалами. Тут був використаний новий підхід із залученням графових баз даних. Графові бази даних — це метод з’єднання даних за допомогою теорії графів. З’єднання частин даних, які називаються вузлами, із зв’язком між ними дозволяє створити структуру, яку можна використовувати для зберігання, запитів і представлення даних.

У цій роботі досліджувалися унікальні властивості графових баз даних, які відокремлюють їх від різних типів баз даних. Було досліджено, як ці властивості можна використати для розробки системи, яка може рекомендувати виробів на основі виробника матеріалу чи типу задач які має виконувати обладнання. Систему, що використовує базу даних такого типу, ще потрібно впровадити в комерційну систему, тому це область розробки рекомендаційної системи, яка має потенціал для багатьох подальших досліджень.

У цьому дослідженні спеціально розглядається властивість бази даних графів, яка називається кластеризацією. Оскільки дані представлені геометрично, можна знайти фрагменти даних, які згруповані разом у кластери, використовуючи багато різних алгоритмів. Існує гіпотеза, що при кластеризації

даних виробів будуть виявлені групи подібних виробів. Ця робота намагається відповісти на два запитання: чи можна успішно представити базу даних, що містить дані про виробів, як базу даних графів? Чи можна використовувати алгоритми кластеризації графів для пошуку наборів виробів, які можна використовувати як рекомендаційну систему?

Область математики, до якої належать бази даних графів, називається теорією графів. Протягом своєї історії було розроблено багато алгоритмів для кластерних графіків. Незважаючи на те, що алгоритми були розроблені для виявлення кластеризації, наразі немає дослідження здатності цих алгоритмів знаходити кластери, які можна використовувати на даних виробів для створення системи рекомендацій. Насправді, дослідження впливу цих алгоритмів на дані виробів взагалі не проводилося. Поки невідомо, як будуть структуровані кластери, знайдені за допомогою цих алгоритмів.

Хоча графові бази даних використовувалися для розробки рекомендаційних систем, не було опублікованих досліджень щодо використання кластеризації як техніки.

Графові бази даних виявилися дуже корисними в комерційному світі. Всесвітня мережа — це графічна база даних. Розвиток Інтернету в цей граф був важливим для повсякденного використання, пошукова система Google базується на алгоритмі графу, який називається рейтингом сторінок. Використання тут алгоритмів графів було революційним для того, як функціонує суспільство. Зрозуміло, що графові бази даних є потужним інструментом в інформаційному світі, і тому варто дослідити використання баз даних графів, теорії графів і алгоритмів графів з метою розробки рекомендаційної системи.

На поточні рекомендаційні системи, розроблені великими сервісами потокового відео, такими як Amazon та Netflix, було вкладено багато часу та грошей. Наявні системи працюють, рекомендуючи вироби на основі схожості інших попередніх пошуків, це називається спільною фільтрацією. Основна відмінність від дослідженого тут підходу полягає в тому, що рекомендація

ґрунтується виключно на власних уподобаннях користувачів. Ще одна проблема, яку можна вирішити за допомогою поточної системи рекомендацій, полягає в тому, що в міру появи нових виробів базу даних можна негайно оновлювати та повторно кластеризувати, що дозволяє легко рекомендувати їх.

Однак спільна фільтрація повинна дати час, щоб новий виріб побачило достатньо людей, щоб дати значущу рекомендацію. Це називається проблемою холодного пуску. Оскільки це поширена проблема в рекомендаційних системах, важливо, щоб дослідження було проведено для найкращого врахування цього.

1 АНАЛІТИЧНА ЧАСТИНА

1.1. Рекомендаційні системи

Системи рекомендацій отримали широке визнання та викликали підвищений інтерес громадськості протягом останнього десятиліття, прокладаючи шлях для нових можливостей продажу в інтернет комерції. Наприклад, інтернет-магазини, такі як Amazon.com, успішно використовують широкий спектр різних типів рекомендаційних систем.

Їх головна мета полягає в тому, щоб зменшити складність для людини, профільтрувати дуже великі набори інформації та вибрати ті фрагменти, які є релевантними для активного користувача. Крім того, рекомендаційні системи застосовують методи персоналізації, враховуючи, що різні користувачі мають різні переваги та різні інформаційні потреби. Наприклад, якщо припустити, що це сфера рекомендацій щодо книг, то історики нібито більше зацікавляться середньовічною прозою, наприклад, Кентерберійськими оповіданнями Джефрі Чосера, ніж літературою про самоорганізацію, яка може бути більш актуальною для дослідників штучного інтелекту.

1.1.1 Збір інформації про переваги

Таким чином, щоб генерувати персоналізовані рекомендації, адаптовані до конкретних потреб активного користувача, системи рекомендацій повинні збирати інформацію про особисті переваги, наприклад, історію покупок користувача, дані кліків, демографічну інформацію, тощо. Традиційно вирази переваг користувачів ai, щодо продуктів bk називають рейтингами ri(bk). Розрізняють два різних типи рейтингів:

Відверті оціни. Користувачі зобов'язані чітко вказати свої переваги щодо будь-якого конкретного предмета, зазвичай, вказуючи ступінь оцінки за 5-бальною або 7-бальною шкалою лайкерта. Потім ці шкали перетворюються на числові значення, наприклад безперервні діапазони $ri(bk) \in [-1, +1]$.

Негативні значення зазвичай вказують на неприязнь, тоді як позитивні значення виражають симпатію користувача.

Неявні оцінки. Відверті оцінки доставляють користувачам додаткові зусилля. Таким чином, користувачі часто намагаються уникати тагаря прямого вказання своїх уподобань і або залишають систему, або покладаються на “те що буде”. З іншого боку, отримання інформації про переваги з простих спостережень за поведінкою користувачів є менш нав’язливим. Типовим прикладом неявних рейтингів є дані про покупку, час читання новин та поведінка при перегляді. Незважаючи на те, що їх легше зібрати, неявні рейтинги мають серйозні наслідки. Наприклад, деякі покупки є подарунками і, таким чином, не відображають інтереси активного користувача. Більше того, висновок про те, що покупка означає симпатію, не завжди справедливий.

Через труднощі з отриманням чітких рейтингів, деякі постачальники послуг з рекомендаційних товарів використовують двосторонні підходи. Наприклад Amazon.com вичисляє рекомендації на основі явних оцінок, коли це можливо. У разі недоступності замість них використовуються спостережувані неявні оцінки.

1.1.2 Типи та методи рекомендаційної системи

Виникли дві основні парадигми для обчислювальних рекомендацій, а саме фільтрація на основі вмісту та спільна фільтрація, або колаборативна фільтрація. Фільтрація на основі вмісту, яку також називають когнітивною фільтрацією, обчислює подібність між кошиком цінних продуктів активного користувача аі та продуктами зі всього додатку продуктів, які досі невідомі для аі. Схожість продукту визначається за допомогою вибраних атрибутів на властивостями цього продукту. Тоді як спільна фільтрація, яку також називають соціальною фільтрацією, обчислює схожість між користувачами на основі їх рейтингового профілю. Більшість подібних користувачів потім служать “порадниками”, пропонуючи найбльш релевантні продукти активному користувачеві.

Розширені системи рекомендацій, як правило, поєднують спільну фільтрацію та фільтрацію на основі вмісту, намагаючись пом’якшини

недоліки будь-якого підходу та використовувати синергетичні ефекти. Ці системи отримали назву “гібридні систем”.

1.2. Гібридні рекомендаційні системи

Гібридні підходи спрямовані на уніфікацію спільної фільтрації та фільтрації на основі вмісту в рамках однієї системи, використовуючи синергетичні ефекти та пом'якшуючи недоліки, притаманні кожній із парадигм. Отже, гібридні рекомендації використовують як інформацію про рейтинг продукту, так і описові характеристики. Насправді можна уявити безліч способів поєднання аспектів співпраці та змісту, Берк перераховує безліч методів гібридизації. Найбільш широко поширеною серед них, однак, є так звана парадигма “співпраці через контент”, коли профілі на основі вмісту створюються для виявлення подібності між користувачами.

Однією з найперших гібридних рекомендаційних систем є Fab, яка пропонує своїм користувачам веб-сторінки. Melville та Hayes and Cunningham використовують інформацію про вміст для прискорення процесу спільної фільтрації. Торрес пропонує різні гібридні системи для рекомендації цитування наукових робіт. Хуанг використовує функції на основі вмісту, щоб побудувати графіки кореляції для дослідження транзитивних асоціацій між користувачами.

Модельно-керовані гібридні підходи ули запропоновані Basilico та Hofmann пропонуючи навчання персептрону та функції ядра, а також Шейн використовуючи більш традиційні байєсівські класифікатори.

1.3. Оцінка рекомендаційних систем

Оцінки рекомендаційних систем є незамінними для того, щоб кількісно оцінити, наскільки корисні рекомендації, зроблені системою Sx, порівнюються з Sy для повного набору користувачів A. Онлайн оцінки, тобто пряме запитання користувачів щодо їхньої думки, у більшості випадків не являється вибором.

Розгортання. Для виконання онлайн-оцінок необхідна цілісна віртуальна спільнота, здатна запускати служби рекомендаційної системи. З іншого боку, успішне розгортання онлайн-спільноти та забезпечення її самопідтримки є громіздким і може вийти за межі більшості дослідницьких проектів.

Нав’язливість. Навіть якщо онлайн-спільнота є легкодоступною, оцінювання не можна просто виконувати за бажанням. Багато користувачів можуть розглядати анкети як додатковий тягар, що не дає їм негайної винагороди, і, можливо, навіть вирішить залишити систему.

Отже, дослідження в основному спиралися на офлайнові методи оцінки, які застосовні до наборів даних, що містять попередні рейтинги продуктів, таких як, наприклад, добре відомі набори даних MovieLens і Every Movie, обидва загальнодоступні. Методи машинного навчання які називаються крос валідація застосовуються до цих наборів даних, наприклад, утримання, К-згортання, або тестування без урахування, а також для метрики оцінки. У наступних розділах наведено опис популярних показників, які використовуються для оцінки в автономному режимі. Розширене та більш повне опитування надано Herlocker.

1.3.1 Показники точності

Показники точності були визначені насамперед для двох основних завдань: по-перше, щоб судити про точність окреми передбачень, тобто наскільки прогнози wi(bk) для продуктів bk відхиляються від фактичних рейтингів ai ri(bk). Ці показники особливо підходять для завдань, де передбачення відображаються разом із продуктом, наприклад анотація в контексті. По-друге, метрики підтримки прийняття рішень оцінюють ефективність допомоги користувачам у виборі високоякісних елементів із набору всіх продуктів, загалом припускаючи бінарні переваги.

1.3.2 Показники точності прогнозування

Показники точності прогнозування визначають, наскільки близькі прогнозовані оцінки до реальних оцінок користувачів. Найбільш відомі та широко використовувані, середня абсолютна помилка (MAE) являє собою

ефективний засіб для вимірювання статистичної точності прогнозів wi(bk) для множин продуктів:

$$|\overline{E}| = \frac{\sum_{b_k \in B_i} |r_i(b_k) - \omega_i(b_k)|}{|B_i|}$$

Рис 1.8 Середня абсолютна помилка

Пов’язана з MAE, середня квадратична помилка (MSE) квадратує помилку перед підсумовуванням. Таким чином, великі помилки стають набагато виразнішими, ніж малі. Дуже прості в реалізації показники точності прогнозування не підходять для оцінки якості топ-N списків рекомендацій: користувачів хвилюють лише помилки для продуктів високого рангу. З іншого боку, помилки передбачення для продуктів низького рангу не мають значення, оскільки користувач все одно не цікавиться ними. Однак MAE і MSE враховують обидва типи помилок абсолютно одинаково.

1.3.3 Показники підтримки рішень

Точність і відкликання, добре відомі з пошуку інформації, не враховують прогнози та їх відхилення від реальних рейтингів. Вони радше судять про те, наскільки актуальними для активного користувача є набір рейтингових рекомендацій. Як правило, перед використанням цих показників застосовується K-згортання, поділяючи оцінені продукти кожного користувача $a_i\ b_k \in Ri = \{b \in B | ri(b) = \not\exists\ 1\}$ на K неперетинних фрагментів бажано однакового розміру. Зазвичай припускаються параметри складання $K \in \{4,5,\ldots,10\}$. Далі K-1 випадково вибраних зрізів використовується для формування навчальної множини Rxi для ai. Ці рейтинги потім визначають профіль Ai, на основі якого обчислюються остаточні рекомендації. Для використовується для передбачення. Цей зріз позначений Txi, становить тестовий набір, тобто ті продукти, які планують передбачити рекомендаційні алгоритми. Точність, Відкликання та F1 Сарвар представляє адаптований

варіант відкликання, фіксуючи відсоток продуктів тестового набору $b \in Txi$, що зустрічаються в списку рекомендацій Pxi щодо загальної кількості продуктів тестового набору $|Txi|$.

$$\text{Відкликання} = 100 \cdot \frac{|T_i^x \cap \Im P_i^x|}{|T_i^x|}$$

Рис 1.9 Визначення відкликання

Символ $\Im Pxi$ позначає зображення карти Pxi , тобто всі елементи списку рекомендацій. Відповідно точність представляє відсоток продуктів тестового набору $b \in Txi$, що зустрічаються в Pxi , по відношенню до розміру списку рекомендацій.

Інший популярний показник, який широко використовується в дослідженнях систем пошуку інформації та рекомендацій є стандартною метрикою поєднує точність і відкликання в одній метриці, надаючи їм однакову вагу:

$$F1 = \frac{2 \cdot \text{Відкликання} \cdot \text{Точність}}{\text{Відкликання} + \text{Точність}}$$

Рис 1.10 Розрахунок точності і відкликання

Оцінка Бріз вводить цікаве розширення для відкликання, відоме як зважене відкликання або оцінка Бріза. Основна ідея відноситься до інтуїції, що очікувана корисність списку рекомендацій - це просто ймовірність перегляду рекомендованого продукту, який насправді є релевантним, тобто взятий з набору тестів, помножених на його корисність, яка дорівнює 0 або 1 для неявних оцінок. Крім того, Бріз стверджує, що кожен наступний елемент у списку менш ймовірно буде переглянутий активним користувачем з експоненційним спадом.

$$H(P_i^x, T_i^x) = \sum_{b \in (T_i^x \cap \Im P_i^x)} \frac{1}{2^{(P_i^{x\,-1}(b)-1/(a-1)}}$$

Рис 1.11 Очікувана корисність рейтингового списку

Параметр a позначає період напіврозпаду перегляду. Період напіврозпаду - це кількість продуктів в списку, при якій існує 50% ймовірність того, що активний агент, представлений навчальним наобром Rxi, перегляне цей продукт. Нарешті, зважене відкликання Pxi щодо Txi визначається наступним чином:

$$BScore(P_i^x, T_i^x) = 100 \cdot \frac{H(P_i^x, T_i^x)}{\sum_{k=1}^{|T_i^x|} \frac{1}{2^{(k-1)/(a-1)}}}$$

Рис 1.12 Розрахунок зваженого відкрикання

Цікаво, що при припущенні $a = \infty$ оцінка Бріза ідентична незваженому відкликанню. Інші популярні показники підтримки прийняття рішень включаються ROC, так звана робоча характеристика приймача. ROC вимірює ступінь, до якого система фільтрації інформації здатна успішно розрізняти сигнал і шум. Менш части використовуваний NDPM порівнює два різних, слабко впорядкованих рейтинга.

1.3.4 За межами точності

Хоча показники точності є важливим аспектом корисності, є риси задоволеності користувачів, які вони не можуть охопити. Тим не менш, показники, що не є точними, до цього часу в основному не викликали серйозного дослідницького інтересу, і вони розглядалися лише як незначно важливі доповнення до показників точності.

1.3.5 Покриття

Серед усіх показників оцінки неточності, покриття було найбільш часто використовуваним. Покриття вимірює відсоток елементів у проблемній області, для яких можна зробити прогноз.

Наприклад, припустивши підхід спільної фільтрації на основі користувачів, представлений у розділі 1.3.2.2.1, охоплення для всієї групи користувачів обчислюється наступним чином:

$$\text{Покриття} = 100 \cdot \frac{\sum_{a_j \in A} |\{b \in B \mid \exists\, a_j \in prox(a_i) : r_j(b) \neq \perp\}|}{|B| \cdot |A|}$$

Рис 1.1 Розрахунок покриття наданих груп користувачів

1.3.6 Новизна і випадковість

Деякі рекомендації дають дуже точні результати, які все ще марні на практиці, наприклад, пропонують банани покупцям у продуктовому магазині: майже всі цінують банани, тому їхні рекомендації мають на увазі високу точність. З іншого боку, через їх високу популярність більшість людей інтуїтивно купують банани, зайшовши в продуктовий магазин. Вони не потребують додаткової рекомендації, оскільки вони “вже знають”.

Таким чином показники новизни та випадковості вимірюють неочевидність зроблених рекомендацій, покарання за “збір вишні”.

2 ТЕХНОЛОГІЧНА ЧАСТИНА

2.1. Графова база даних

Графова база даних - це NoSQL база даних, яка має структуру графу. Вона складається з вузлів, які представляють сутності даних які з’єднані між собою ребрами, які представляють відносини між цими об’єктами. Графові бази даних надзвичайно корисні для сильно взаємопов’язаних даних і як такі викликали великий інтерес у галузях біології, інформатики та інформаційних технологій. Юн та співавтори, у статті 2017 року “Використання графових баз даних для інтеграції гетерогенних біологічних даних”, успіх графових баз даних у складних біологічних проблемах. Використання цієї технології в такій складній задачі показує силу, яку мають ці бази даних. Це стаття з великим цитуванням, яка показує справжній потенціал цієї технології.

Графова база даних показана на рисунку 2.1. Цей рисунок було взято з офіційного сайту Neo4j, велика компанія програмного забезпечення для графічних баз даних. Оскільки ця компанія є основним джерелом технології графових баз даних, вона є і корисним джерелом інформації про графові бази даних.

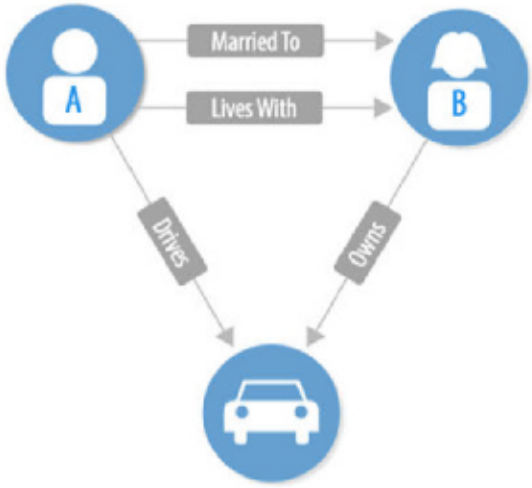

Рис. 2.1. Приклад графової бази даних

На рисунку 2.1 показано як дві людини A і B, пов’язані один з одним, і як обидва пов’язані з автомобілем. Перша пропозиція щодо графової бази даних як методу опису даних була в статті Джона Сови 1979 року “Концептуальні графіки для інтерфейсу бази даних”. У цій статті концепція графічної бази даних розглядається як засіб опису даних, а не як зберігання. Граф мав бути посередником між людиною-користувачем і комп’ютером. Концепція була задумана як метод запиту комп’ютерної системи. Система повинна була перетворити питання з людської мови на концептуальний графік. Система може шукати інші графи в базі даних, які мають відношення до вихідного запитання.

З тих пір багато систем графових баз даних було розроблено комерційно. Першою комерційною графовою базою даних була Allegro graph у 2004 році, ця база даних спочатку була розроблена для зберігання трійок RDF. Дуже помітним прикладом використання графової бази дани є Amazon Neptune, вперше обговорений у 2011 році в статті Кріса Бранча “Neptune: мова, специфічна для домену для розгортання програмного забезпечення на хмарних платформах”. Система була випущена в 2018 році. Це розширення веб-сервісів Amazon. Amazon стверджує, що ця система є швидшим і ефективнішим способом запуску веб-додатків, які працюють з неймовірно великими наборами даних. Хоча ця інформація взята з документації Amazon з невеликими доказами. Через те, що система нова, часу на створення повних світів про її можливості було мало. Багато інформації про Neptune є конфіденційною, оскільки цей продукт є великою інвестицією для Amazon.

Також було випущено багато популярних систем графічних баз даних з відкритим вихідним кодом. Одним із популярних прикладів є Neo4j, програма запитів до графової бази даних на основі Java, .NET, JavaScript, Python, Ruby, яка була випущена в 2007 році. Ця програма може бути використана для зберігання даних у форматі графа.

Також були розроблені мови запитів, щоб збільшити корисність графових баз даних, як методів зберігання даних. Поширеною мовою є Cypher.

2.2. Графові бази даних для рекомендаційних систем

У цій роботі було дослідження графове представлення як потенційний метод рекомендації для цифрових бібліотек. Було створено мережу графів між двома базами даних SQL. Перший рівень - це вміст книги, а другий - демографічні дані клієнтів. Як і у стандартних графових базах даних, точки даних розглядалися як вузли, а відносини як ребра. Однах з'єднання вузлів могло відбуватися лише між двома шарами. Навіть з цим обмеженням, при оцінці людей було отримано похибку 18.3% і точність 38.1%. Хоча предметна оцінка цього дослідження була неймовірно обмеженою. Суб'єктами були лише два студенти.

Лише в середині кінця 2000-х років були створені справжні комерційні графові бази даних з такими пакетами як neo4j у 2010 році. Ефективність системи графової бази даних на відміну від стандартної реляційної бази даних була продемонстрована в статті "Chad Vicknair et al 2010 A Comparison". У статті досліджуються два типи запитів: структурні запити та запити даних. Запити даних: підрахувати кількість вузлів, дани корисного навантаження яких дорівнюють деякому значенню. Підрахувати кількість вузлів, дані корисного навантаження яких менші за деяке значення. Підрахувати кількість вузлів, дані корисного навантаження яких містять певний рядок пошуку (довжина коливається від 4 до 8). Результатом цього було те що Neo4j працював краще ніж MySQL, виконуючи запити швидше. Однак це не стосується конкретно рекомендаційних систем.

На початку 2010-х років було розроблено кілька мов запитів до графічних баз даних. У 2013 році Holzschuher. порівняли мови графічних баз даних Cypher, Gremlin та Neo4j у своїй роботі 2003 року "Продуктивність Графових мов запиту". Дані тесту містили 2011 осіб, 26982 повідомлення, 25365 видів діяльності, 2000 адрес, 200 груп і 100 організацій. Це значна кількість даних, яка повинна стати хорошим тестом мови на основі графів. Отриманий результат показав, що пови графових баз даних були на порядок

вищі ніж мови реляційних баз даних. Стосовно рекомендаційних систем, Frano is Fouss розробили спільну роботу на основі графових баз даних, “Експериментальне дослідження ядер графів на задачі спільної рекомендації”.

У цій статті описано створення системи рекомендацій на основі даних з загальнодоступних джерел. Ці дані складаються з виробників, виробів і матеріалів. Вузли цього графа були визначені людьми а ребра - відносинами. А саме “material”, між предметом і матеріалом, та “belongs_to” між матеріалом та категорією. Це викликало кластеризацію вузлів у межах гарфа. Людям в кластері будуть рекомендовані матеріали, які використовувалися раніше при схожому запиті. Кластери визначають шляхом розрахунку евклідової відстані між вузлами. Результат цієї роботи показав, “що три заходи подібності забезпечують хорошу та стабільну роботу”.

Очевидна перевага графовик баз даних полягає в тому, що, оскільки рекомендація не базується на інформації споживачів, на неї не впливає проблема холодного запуску.

2.3. Кластеризація в теорії графів

Хоча єдиного визначення кластера в наборі даних графі немає, Шеффер визначає кластер як “дані, такі, що елементи, присвоєні конкретному кластеру, подібні або пов’язані в певному, заздалегідь визначеному сенсі”. Бар-Ілан та ін. визначили набір вимог, яким має відповідати набір даних, щоб бути прийнятим як кластер, визначений шляхами на графі. У теорії графів шлях - це послідовність ребер, що починається з вузла N0 і закінчується у вузлі Nk. Визначення Бар-Ілана таке:

- Кожен кластер повинен бути інтуїтивно підключений.
- Повинен бути принаймні один, а краще декілька шляхи, що з’єднують кожну пару вершин кластеру.
- Шляхи повинні бути підключені всередині кластеру.
- Два вузли в кластері повинні бути не тільки по шляху, який проходить через них, але і по шляху, який відвідує лише вузли в кластері. Дж. М. Клейнберг визначає кластер з точки зору його щільності, відношення

наявних ребер до максимальної кількості можливих ребер.Клейнбер вважає “хорошим” кластером, де підграф який утворює кластер, є щільним, але має відносно мало зв’язків з вузлами кластера до решти графі.Це визначення спирається на досить туманну термінологію.

Проблема з цими теоретичними визначеннями кластерів полягає в тому, що вони непридатні до проблем реального світу. Однак існують алгоритмічні способи оцінки кластера. П’ять поширених підходів до виявлення кластерів. які обговорюються в книзі М. Нідхема та А.Е. Ходлера - це кількість трикутників і коефіцієнт щільності, міцно з’єднані компоненти, поширення етикетки та модульність Лувена.

Вони визначають вимірювання кількості трикутників, скільки вузлів утворює трикутники і ступінь до якої вузли мають тенденцію до групування. Сильно зв’язані компоненти як алгоритм, який знаходить групи, де кожен вузол доступний з кожного іншого взула в тій самій групі, дотримуючись напрямку зв’язку. Поширення міток як алгоритм, який визначає кластери, розповсюджуючи мітки на основі більшості околиць, і модульність Лувена як алгоритм, який максимізує передбачувану точність групування шляхом порівняння ваг і щільностей відносин із визначеною оцінкою або середнім значенням. Потім у книзі показано, що алгоритм сильно зв’язаних компонентів ідеально підходить для рекомендаційних систем. Однак визначення алгоритму надзвичайно високого рівня і немає експериментальних доказів, які б підтвердили це твердження.

Т. Шанканд Д. Вагнер провів ретельний аналіз кількості трикутників і коефіцієнта щільності у своїй статті 2005 року “Finding, Counting and Listing all Triangles in Large Graphs, An Experimental Study”. Вони визначили трикутник як тривузловий підграф і дослідили кілька методів підрахункі трикутників. Успіх алгоритмів підрахунку трикутників вимірювався часом виконання та кількістю операцій з трикутником. Визначення операцій трикутника варіювалося між алгоритмами, але по суті було асимптотичним часом виконання.

Алгоритми були запущені в мережі з не орієнтованими графами, мережею доріг Німеччини та орієнтованим графом. Для мереж біт метод ітерації вузла був найшвидшим у виконанні, він був найбільш асимптотично інтенсивним. Цей метод покладався на ітерацію кожного вузла та підрахунок навколишніх ребер для підрахунку кожного трикутника. Алгоритм із прямим хешуванням був найгіршим. Натомість цей алгоритм працював шляхом ітерації динамічних даних.

Алгоритм сильно пов’язаних компонентів був одним із перших алгоритмів виявлення скупчення. Цей алгоритм був винайдений Робертом Тарьяном у 1970 році. Алгоритм працює шляхом пошуку наборів вузлів, де всі вузли можуть бути досягнуті всіма іншими вузлами в обох напрямках, але не обов’язково безпосередньо. Еско Нуутіла та Ельяс Сойсалон-Сойнінен покращили цей алгоритм у 1990 році, що робить його здатним обробляти розріджені графи та тривіально компоненти.

Модульність Лувена - алгоритм кластеризації графів, розроблений у 2008 році Вінсентом. Д. Бонделем та співавторами на основі розробки модульності в системі. Модульність - це метод кількісного визначення міцності кластера в системі графів. Модульність - це значення від -1 до 1, яке вимірює щільність ребер всередині кластера графів порівняно з щільністю ребер за межами цього кластера. Математичне визначення модульності показано нижче.

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

Рис. 2.2. Математичне визначення модульності

Де Aij - вага між вузлами i та j, ki та kj - це сума ваг вузлів, приєднаних до і та j відповідно. 2m - це сума всіх ваг ребер у графі, сі та cj - спільноти вузла, а δ - проста дельта функція.

Модульність була важливим компонентом у багатьох областях, які можна представити в мережі графів. Це включає всесвітню павутину,

метаболічні мережі, соціальні моделі, і тому алгоритм кластеризації, заснований на модульності, був важливою віхою для досягнення.

Алгоритм модульності Лувена є двоетапним ітераційним процесом, який описано в рівнянні нижче.

$$\Delta Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2m} - \frac{\sum_{tot} + k_i^2}{2m}\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

Рис. 2.2. Алгоритм модульності Лувена

$i \quad n$

Де $\sum$ - це сума всіх ваг зв’язків всередині спільноти, у яку вона переходить, $\sum$ - це сума всіх ваг посилань на вузли спільноти. З точки зору рекомендаційних систем, модульність Лувена була використана для створення систем рекомендацій виробів. У 2015 році Діпіка Лалвані та інші опублікували роботу, в якій використовували модульність Лувена для пошуку кластерів у графовій базі даних соціальних мереж для пошуку та рекомендацій. В іншому дослідженні, проведеному в 2013 році, використали дані IMDB для побудови графічної бази даних, яка описувала загальні зв'язки між людьми. Хоча обидва експерименти дали точні рекомендації з кластера, обидва залежать від інформації користувачів. Поки що не проводилось розслідування щодо використання модульності Лувена в базі даних графів, яка містить лише інформацію про певні записи.

Поширення міток- це система, яка працює шляхом позначення та перемаркування вузлів у графі до тих пір, поки не буде створено задовільний кластер міток. Ця техніка знайшла застосування в медичній сфері. Рольф А. Хекерманн та співавтори використовували поширення міток на МРТ-сканування мозку, щоб досягти більшої автоматичної анатомічної МРТ сегментації мозку, поєднуючи поширення мітки та злиття рішень з позитивними результатами.

У рамках математики були розроблені концепції для розв’язування складних задач неоднорідного бігармонічного рівняння з граничними умовами Діріхле. Це обговорювалося в статті Джіндонг Ванг “Лінійне поширення сусідства та його застосування”. Однак реальне застосування цього алгоритму до складних математичних задач ще не повністю реалізовано.

Зх Ву та співавтори довели використання поширення міток у соціальних мережах для пошуку спільнот, що перекриваються, у свіїй роботі 2012 року “Збалансоване поширення кількох міток для виявлення перекриваючих спільнот у соціальних мережах”. Поширення міток також виявилося корисним для стиснення графів, як було показано у 2011 році Паоло Болді та інших у своїй статті “Розповсюдження міток по шарах”.

Однак з точки зору рекомендаційних систем, Хоу Кіанг та інші довели потенціал його використання, поєднавши його зі спільною фільтрацією. Це обговорювалося в їхній статті 2012 року “Метод персоналізованої рекомендації, заснованої на розповсюдженні кількох міток для виявлення спільнот, що перекриваються”. У цій статті було досліджено набір даних MovieLens і виявлено ледь покращені результати щодо базового порівняння спільної фільтрації. У межах музични рекомендацій Бо Шао та інші використовували поширення міток для розробки системи рекомендацій, заснованої на графовій базі даних, що описує моделі доступу користувачів до музичної мережі NewWisdom,а також акустичні особливості пісень, які слухав користувач. Це було написано в їхній статті 2009 року “Рекомендації щодо музики на основі акустичних характеристик та шаблонів доступу користувачів”.

2.4. Проектування графової бази даних

Основними цілями цього проекту були, побудова графової бази даних з описом устаткування та матеріалів, аналіз методів кластеризації графів у застосуванні до зазначеної бази даних та дослідження методів кластеризації як методу рекомендації матеріалу фахівцям, на основі опису їх процесу виробництва. Таким чином методологія була розділена на три розділи,

створюючи та досліджуючи різні бази даних, виконання методів кластеризації в базах даних і використання знайдених кластерів для створення рекомендаційної системи.

Було протестовано різні види графових баз даних, щоб визначити оптимальний граф. Хоча граф може бути побудований правильно, з чіткими вузлами та зв’язками, функціональність графу можна дослідити лише шляхом аналізу результатів алгоритму кластеризації. Так само оптимальний метод кластеризації можна знайти лише шляхом порівняння результатів рекомендаційної системи.

2.4.1 Проектування графової бази даних

Було досліджено два підходи до побудови бази даних графів. Перший передбачав створення вузлів для кожного запису та кожної властивості. Тобто для кожного запису є вузол із міткою “ID”, а для кожного іншого поля в базі є вузол “NAME”, “MATERIAL”. Краї між цими вузлами будуть спрямовані відношеннями від вузла властивостей до вузла виробу. Основа цього графа полягає в тому, що вузли властивостей будуть групуватися з вузлами записів, після чого записи можуть бути ідентифіковані з цих кластерів під час рекомендації.

Другий підхід полягав у побудові графа, який містив лише вузли виробу. Вузли виробу потім можуть бути з’єднані двонаправлено з ребрами, які представляють спільні властивості. Наприклад, якщо у виробу є один із тих самих матеріалів, буде створено двонаправлений зв’язок із позначкою “BASE MATARIAL”. Те ж саме було зроблено для інших розглянутих об’єктів. Очікувалося, що за допомогою графіків, розроблених за цим методом, кластери будуть формуватися на основі зв’язків між кожним записом, а не на основі вузлів властивостей, спільних між записами. Оскільки кластери будуть повністю складатись з записів, весь кластер можна використати для рекомендацій.

2.4.2 Використані дані

Даними, використаними для проектування графової бази даних, була база даних TMDb 5000 Iron. TMDb (The Iron Database) - це онлайн база даних про різні типи металів, створена спільнотою. Веб сайт збирає дані з 2008 року. Хоча публічно розміщені та модеровані дані, як правило, не такі найдіні, як опубліковані дані. Обсяг веб-сайту передбачає, що їх можна використовувати для аналізу. База даних TMDb 5000 - це вибірка з багатьох матеріалів із добіркою інформації про них.

Дані надійшли у вигляді двох файлів .csv, tmdb_5000_produc.csv і tmdb_5000_credits.csv. Перший містить детальну інформацію про кожен матеріал. Стовпці даних tmdb_5000_produc.csv містять детальну інформацію про про різні властивості того чи іншого матеріалу, ідентифікатор (унікальний для кожного матеріалу).

Дані взяти з онлайн-спільноти дослідників даних. Безперервні дані непридатні для графової бази даних, оскільки їх е можна використовувати для створення окремих, зручних для використання вузлів. Поточна назва була використана замість оригінальну, щоб зберегти узгодженість з іншими даними. Цей CSV мав один рядок для кожного матеріалу які були представлені у форматі JSON.

2.4.3 Використані технології

Для створення графу використовувалися дві технології Python і Neo4j. Python використовується для підготовки даних TMDb 5000 у формат, який можна було використовувати для перетворення шуканих даних у граф. Neo4j було використано для створення графу.

У Python для керування файлами csv використовується бібліотека pandas. Ця бібілотека може імпортувати файли csv та зберегти їх як об’єкт фрейму даних. Це дозволяє легко маніпулювати даними, які розглядаються як стандартна таблиця. Бібліотека здатна читати та зберігати рядки JSON як об’єкти фрейму даних. Це дозволило отримати доступ до властивостей,

описаних як рядки JSON. Бібліотеку також можна використовувати для об’єднання таблиць, що робить її корисною для роботи з даними які розділені на кілька файлів.

Neo4j - це найпоширеніша у світі система керування графовими базами даних. Система працює на основі теорії графів, створюючи вузли, які містять дані, з’єднані ребрами ,які представляють деяке з’єднання даних. Neo4j дозволяє вільно будувати графи, включаючи створення нових вузлів і ребер. Коли вузол створюється, йому потрібно призначити мітку. Це встановлює відмінність між вузлами та можливістю запитувати граф. Вузол з міткою n позначається як (n). Окремі вузли можна відрізними один від одного, призначивши їм властивості під час їх створення. Це дозволяє робити запити до конкретних вузлів на графі. Кількість властивостей на один вузол необмежена.

Властивості можуть бути цілими, плаваючими, рядковими, булевими або масивом будь якого з типів перерахованих вище.

У Neo4j ребра в графі називаються відносинами. Ці відносини можуть бути ненаправленими, однонаправленими. Двонаправлені відношення утворюються двома однонаправленими відношеннями, протилежними один від одного. Як і у випадку з вузлами, зв’язкам необхідно присвоїти мітку. Відношення r між двома вузлами n і m представлено в Neo4j як: (n) - [r] -> (m).

Стрілка не є обов’язкова і вказує напрямок зв’язку. Відносинам можна призначити вагові коефіцієнти подібно до властивостей вузла. Це використовується для побудови зважених графів, які надають значення конкретним відносинам. Коли ваги призначаються, вони повинні бути позначені для правильного запиту.

Перевагою Neo4j є його здатність запитувати базу даних. Програма використовує мову запитів під назвою Cypher для дослідження властивостей графу. Її можна використовувати для пошуку конкретних даних, побудови конкретних підграфів і застосування таких операцій, як підрахунок чи середнє. Вона також дозволяє маніпулювати властивостями вузла за допомогою

числових і рядкових операцій.

Запити в Cypher орієнтовані навколо двох команд: MATCH і RETURN. MATCH використовується для пошуку, RETURN для повернення результату. Між цим до певних вузлів можна отримати доступ і маніпулювати ними за допомогою однієї з їхніх властивостей. Всередині запиту можна виконувати логічні, рядкові та числові операції так само як SQL. Для більш просунутих статистичних і математичних операцій необхідні пакети в Neo4j.

Існують пакети, які дозволяють проводити більш складний аналіз. Два пакети, використані в цьому проекті, були APOC (Awesome Procedures on Cypher) і Graph Algorithms. APOC містить багато корисних функцій і особливо корисний для виконання різних статистичних рівнянь. Graph Algorithms використовується для застосування алгоритмів специфічних для теорії графів. Вони включають централізованість, пошук шляху, передбачення зв’язків і виявлення спільноти.

2.4.4 Граф з багатьма мітками

Перший досліджений граф складався з вузлів з декількома мітками. Кожен прилад був позначений його назвою та ідентифікатором. Кожна властивість цього приладу також було створена як вузол із відповідною міткою. Наприклад, вузли GENRE і ACTOR.

Щоб звузити граф з імпортованих даних у Neo4j, він повинен бути в одному csv. Коли csv зчитується в таблиці, можна перебирати рядки, а значення використовувати в побудові вузлів, зв’язків і властивостей. Оскільки дані були у формі двох файлів csv, їх потрібно було об’єднати в один таким чином, щоб за допомогою ітерації можна було побудувати потрібний граф.

Для графа з кількома мітками таблиця повинна мати стовпець для кожного досліджуваного приладу та властивості. З даними в цьому форматі кожен стовпець можна вважати міткою вузла, кожен унікальний запис є вузлом і відносини можна побудувати з кожного рядка.

Щоб це сталося, для приладу повинно бути кілька записів, щоб відповідати кожній із властивостей. Для цього було створено фрейм даних

pandas для кожного файлу csv. Окремі властивості кожного рядка були вбудовані в новий фрейм даних, стовпець для значення та стовпець для ідентифікатора приладу. Потім кадри даних були зовнішньо об’єднані за допомогою ідентифікатора приладу. Останніми стовпцями були ідентифікатор приладу, назва, вартість, дата випуск, ідентифікатор кольору, габаритні розміри, виробник, ідентифікатор виробника.

Були створені різні таблиці з різною кількістю моделей, щоб можна було дослідити вплив кількості модель. Було створено сім таблиць, що містять від одного до семи виробника на прилад.

Потім граф було побудовано за допомогою Neo4j. З попередньої таблиці потрібно було створити чотири типи на вузлі: (manufacturer), (product), (color), (model). Вони містили відповідне ім’я та ідентифікатор. Було побудовано такі відносини:

- (:manufacturer)-[:manufactured_in]->(:product)
- (:color)-[:manufactured]->(:product)
- (:product)-[:is_model]->(:model)

Процес створення графа за допомогою цього програмного забезпечення можна розбити на кілька кроків. По-перше, було накладено обмеження на створення вузлів, щоб вузли manufacturer мали унікальний ідентифікатор виробника, вузли product мали унікальний ідентифікатор виробу, вузли color мали унікальний ідентифікатор кольору і щоб назва типу виробу була унікальною Обмеження були накладені на ідентифікатори, а не на імена, оскільки це відрізняє спільні імена. По-друге, рядки таблиці виконуються з новим вузлом manufacturer для кожного ідентифікатора виробника, цьому вузлу було надано ім’я та інформацію про ідентифікатор. Потім цей вузол був об’днаний у граф. Те саме було запущено для кожного типу вузла.

Щоб створити зв’язки, рядки таблиці були запущені знову. Цього разу функція Neo4j MATCH була використана для пошуку вузла, який збігається з ідентифікатором матеріалу та вузла, який відповідає ідентифікатору виробника і створює відносини manufactured_in між ними. Це повторювалося для model та

is_model.

За допомогою цього методу було створено кілька графів для різної кількості таблиць виробників та версії графі з моделлю та без. Це дозволило перевірити вплив кожного параметру.

3 КОНСТРУКТОРСЬКА ЧАСТИНА

3.1. Побудова графу

Запити Cypher успішно побудували граф, що містить вузли manufacturer, вузли model, вузли product і color, а також відповідні зв’язки. Дивлячись на цей рисунок, можна побачити, що потрібні вузли були створені успішно. Також можна побачити, що правильні відносини з'єднують вузол. Тут видно, що всі 4803 вироби в базі даних були додані до графіка, а також усі 20 виробників. Було додано 13 025 матеріалів. Це розумне значення, оскільки зібрано сім матеріалів з кожного виробу, але вузли для кожного матеріалу різні. Отже, значення акторів має бути менше 7.

Існує 4803 вузли product, і відносини material_in були засновані на семи матеріалах у цього продукту. Це означало б, що має бути 33621 спільних відносин. Проте існує лише 32791 стосунків material_in на product. Це в середньому 6,827 матеріалів на виріб. Хоча це дуже близько до прогнозованого, невелика розбіжність свідчить про те, що сталося щось несподіване. Було 650 випадків, коли у виріб було призначено менше семи матеріалів. Ймовірно, це пов’язано з процесом імпорту. Аналогічно, було 4775 зв’язків manufactured_by, тобто 0,994 виробника на прилад. Очікується, що на один виріб буде трохи більше 1 виробника, оскільки всі, крім кількох виробів, мають лише одного виробника. Різниця у відносинах знову дуже незначна і, швидше за все, через брак даних. У зв’язках HAS_GENRE було 2,532 зв’язку на вузол виробу. Якщо порівняти кількість зв’язків із кількістю вузлів, то на вузол model припадає 2,48 відносин ACTED_IN.

3.1.4 Кластеризація

Першим проаналізованим графом був чистий одновузловий граф без доданих вагових показників. Першою мірою було число трикутників і коефіцієнт кластеризації. Для відображення різних результатів було створено діаграму, як показано на рисунку 3.1.

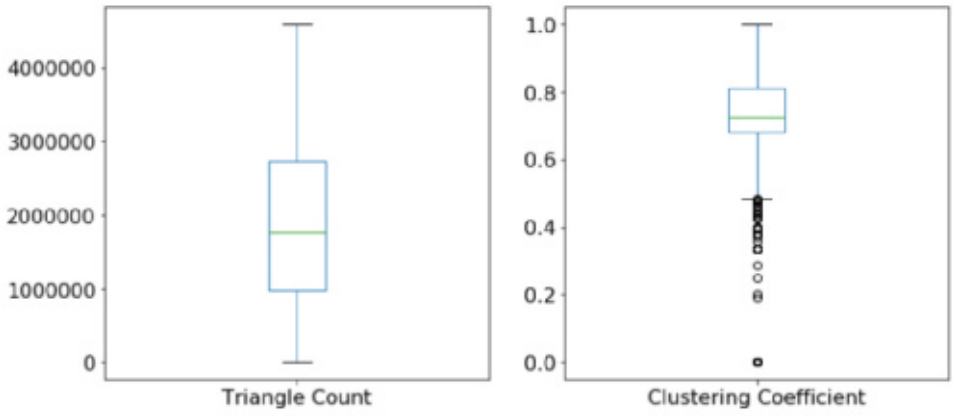

Рис. 3.1. Одновузловий граф.

Дивлячись на кількість трикутників, можна побачити, що існує величезний діапазон від нуля до майже 5000000. Це не дуже хороший знак, оскільки було б бажано, щоб усі вузли містили багато трикутників. Аналогічно, хоча середній коефіцієнт кластеризації 0,786 є скоріше висотою, поширення результатів означає, що багато вузлів не знаходяться в межах відповідного кластера.

Для системи рекомендацій важливо, щоб усі вироби були згруповані в чітко визначені кластери, щоб можна було дати ґрунтовні рекомендації. Другим алгоритмом, запущеним на цьому графіку, був алгоритм сильно зв'язаного компонента. На рисунку 3.2 нижче показано розміри кластерів, знайдених за допомогою алгоритму.

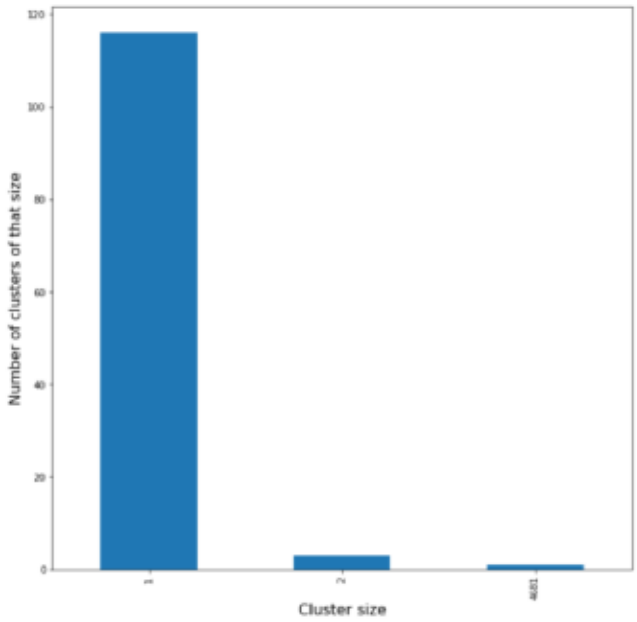

Рис. 3.2. Графік, що показує розміри різних кластерів, а також різну кількість кластерів цього розміру за допомогою алгоритму сильно зв'язаних компонентів.

Цей рисунок показує, що кластеризація розподілена дуже погано. Є один кластер, який містить 4681 із 4803 вироби, решта вписується в кластери розміру 1 або 2. Ймовірно, це пов'язано з щільністю графіка. Оскільки для кожного аспекту виробів існують взаємозв’язки, загальна кількість зв’язків становить 10396650. Це те, що створює таку велику кількість трикутників.

Оскільки трикутників утворено так багато, усі вузли занадто пов’язані. Виконання алгоритму підключеного компонента призвело до тих самих значень, що й для сильно пов’язаних компонентів. Як видно нижче на рисунку 3.3.

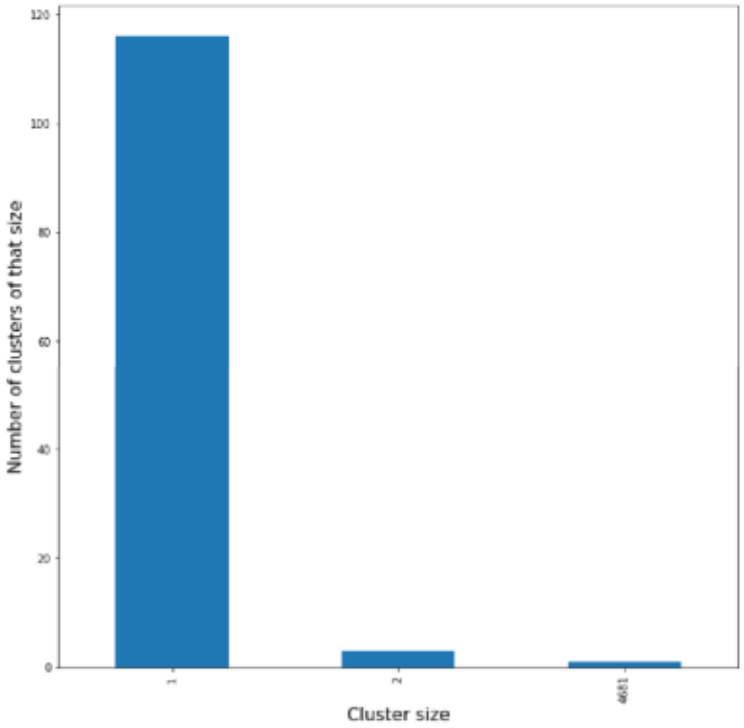

Рис. 3.3. Графік, що показує розміри різних кластерів, а також різну кількість кластерів цього розміру за допомогою алгоритму підключеного компонента.

Це пов’язано з єдиною відмінністю алгоритму в тому, що сильно зв’язані компоненти знаходять групи вузлів, які можна досягти, дотримуючись напряму зв’язку. Алгоритм зв'язаних компонентів, однак, ігнорує напрямок. Ця різниця не впливає на графіки цієї структури, оскільки кожен напрямний вузол між кожним відношенням має інше відношення в протилежному напрямку. Таким чином, вузли будуть доступні незалежно від напрямку зв’язків.

Після запуску алгоритму Лувена на цьому графіку результати були нанесені на стовпчасту діаграму, як показано нижче на малюнку 3.4.

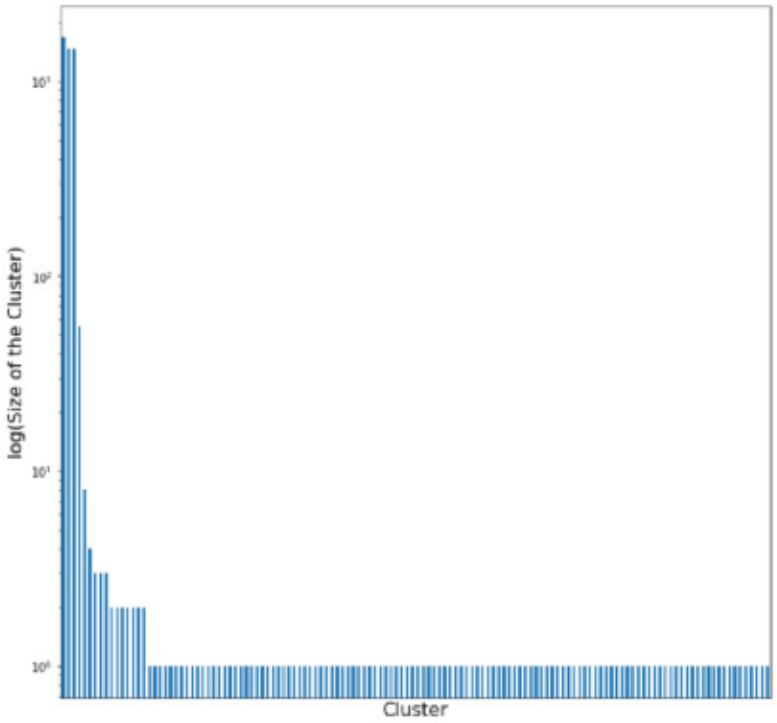

Рис. 3.4. Гістограма, що показує логарифм розміру кожного кластера, виявленого за допомогою алгоритму кластеризації Лувена

На цьому графіку кожен стовпчик представляє кластер, а вісь y — це логарифм 10 базових розмірів. Це було зроблено для кращого порівняння розмірів кожного кластера, враховуючи великі відмінності. У порівнянні з результатами сильно зв’язаного компонента і зв’язаного компонента, кластерів набагато більше. Виявлено 129 кластерів. Більшість виробів об’єднуються в три великі кластери. У ці три кластери було включено 4605 виробів. 116 кластери містили лише один виріб. Сім кластерів містили два вироби, а три — три вироби. Нарешті було три кластери з 4, 8 і 56 виробів. Хоча було виявлено більше кластерів, 89% з них містили лише один вузол. Це означає, що 89% кластерів не можна використовувати для рекомендації виробів. Хоча було більше кластерів із більшою кількістю виробів, цього все одно недостатньо, щоб бути розглянутим для системи рекомендацій.

Причина цього, швидше за все, полягає в кількості зв'язків, що створюють занадто щільний граф. Інший аспект полягає в тому, що алгоритм Лувена в

основному базується на вагових зв'язках. Усі ваги тут були встановлені за замовчуванням 1.0, що зменшує ефект алгоритму. Нарешті, був запущений алгоритм поширення мітки. Щоб не зробити процес занадто складним до обчислень, ітерації були встановлені на 10. Була зроблена ще одна стовпчаста діаграма, яка показувала розміри кластерів, як можна побачити на рисунку 3.5.

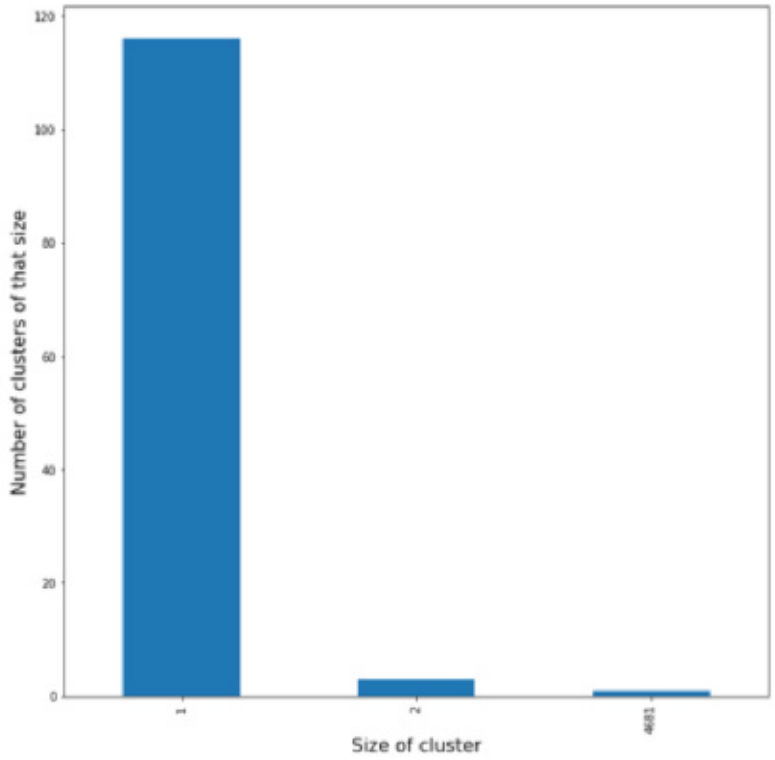

Рис. 3.5. Гістограма, що показує розмір кластерів, знайдених за допомогою алгоритму поширення мітки, і кількість кластерів такого розміру.

Цікаво, що функція поширення мітки знайшла кластери такого ж розміру, як зв’язний компонент, і алгоритми сильно зв’язаних компонентів. Ймовірно, це пов’язано з щільністю графу, а також через відсутність ваг. Це, ймовірно, призвело до того, що алгоритм поширення мітки розповсюдив мітки таким концентрованим способом. Від основного кластера могли бути відокремлені лише вироби, які містили дуже різні властивості.

3.2 Додані ваги

Алгоритми кластеризації були запущені на зваженій версії графу з однією міткою. При дослідженні зваженої версії цього графу не було потреби розглядати алгоритм сильно зв’язного компонента. Це тому, що алгоритм Neo4j не враховує ваги. Таким чином, результат буде таким же. Результати підрахунку трикутників і коефіцієнт кореляції також будуть такими ж, оскільки структура графу не змінюється. Для кожного з запущених алгоритмів було сформовано 4803 кластери. Тобто кожен виріб формується у власний кластер. Додавання ваг в цю структуру графу значно змінило те, як працює кожен з алгоритмів. Коли вага не застосовувався, щільність графіка змушувала алгоритми групувати вироби в кілька дуже великих кластерів.

Введення ваг забезпечило необхідну відмінність. Однак щільність графів все ще перешкоджала показу справжніх шаблонів. Кількість ваг між графом також може мати проблеми під час виконання алгоритму. Якщо дивитися на рисунок 3.2 раніше, діапазон зважень між зв’язками був дуже великим. Це, швидше за все, призвело до більшого поділу. Що стосується модульності Лувена, алгоритм максимізує свою передбачувану точність, порівнюючи щільність з ваговими зв’язками. Щільність зв’язків була дуже високою, а найпоширеніші зв’язки, спільні ключові слова, мали дуже низькі коефіцієнти. Алгоритм не міг побудувати кластери при порівнянні цих двох зв’язків.

Завдяки реструктуризації графіка кількість зв'язків було зменшено. Отже, це зменшило кількість утворених трикутників, а отже, вплинуло на коефіцієнт кластеризації. Графіки нового трикутника та коефіцієнти кластеризації можна побачити на рисунку 3.6.

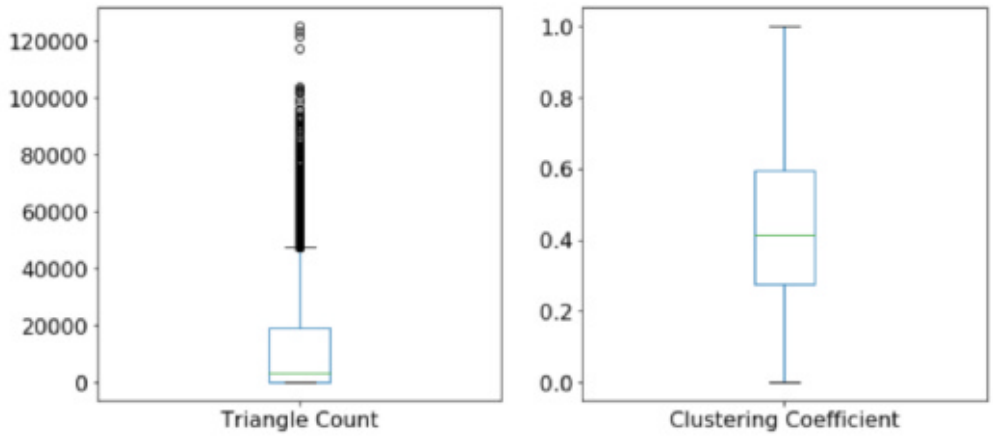

Рис. 3.6. Діаграми кількості трикутників і коефіцієнтів кластеризації.

Як і очікувалося, кількість трикутників було значно зменшено, середнє число трикутників тепер становить 13472,84. В результаті цього середнє значення коефіцієнта кластеризації було в 0,51 рази більше, ніж попередньо агрегований графік. Це помітне зниження здатності багатьох виробів до групування. Першим алгоритмом, запущеним тут, був граф зв'язних компонентів. Різна структура зменшує кількість з'єднань і має призвести до більшої кількості поділів. Нові ваги графіка також були враховані.

За допомогою цього алгоритму було виявлено 678 кластерів, проте 4126 з них були в одному кластері, решта виробів містилися окремо в одному вузлі кластерів. Це далеко не позитивний результат.

4 НАУКОВО-ДОСЛІДНА ЧАСТИНА

4.1 Фільтрація на основі вмісту

Підходи до складання рекомендацій, засновані на змісті, глибоко вкорінені в пошук інформації. Як правило, ці системи використовують байєсівські класифікатори за допомогою функцій змісту, або виконують запити у векторному просторі найближчого сусіда. Байєсівські класифікатори використовують теорему Байєса про умовну незалежність:

$$P(R|F) = \frac{P(F|R) \cdot P(R)}{P(F)}$$

Рис 4.1 Теорема Байєса

Більше того, байєсівські класифікатори роблять “наївне” припущення, що ознаки опису продукту є незалежними, що зазвичай не так. Враховуючи мітку класу, ймовірність незалежності bk до класу Ri враховуючи його n значень ознак $F1,\ldots,F_n$, визначається наступним чином:

$$P(R_i,\ldots,F_n) = \frac{1}{Z} \cdot P(R_i) \cdot \prod_{j=1}^{n} P(F_j|R_i)$$

Рис 4.2 Визначення ймовірності входження продукту до класу продуктів

Змінна Z представляє коефіцієнт масштабування, що залежить лише від $F1,\ldots,F_n$. Ймовірності $P(Ri)$і $P(F_j|R_i)$ можна оцінити за навчальними даними. Для запитів у векторному просторі атрибути, наприклад, терміни простого тексту або машиночитані метадані, витягуються з описів продуктів і використовуються для профілювання користувачів і представлення продукту. Наприклад Фаб представляє документи в термінах 100 слів з найвищою вагою TF-IDF, тобто словами, які зустрічаються частіше в цих документах, ніж у середньому.

4.2. Спільна фільтрація

Фільтрація на основі вмісту працює лише у тих доменах, де вилучення ознак можливе, а інформація про атрибути легкодоступна. Спільна фільтрація, з іншого боку, використовує уявлення без вмісту і не стикається з

тими самими обмеженнями. Наприклад, Джестер, використовує спільну фільтрацію, щоб рекомендувати жарти своїм користувачам. У той час як фільтрація на основі вмісту враховує описові характеристики продуктів, спільна фільтрація використовує оцінки, які користувачі надають продуктам.

Отже, алгоритми колаборативної фільтрації зазвичай працюють з набором користувачів $A = \{a1, a2, \ldots, an\}$, набором продуктів $B = \{b1, b2, \ldots, bm\}$ і функціями часткового оцінювання для кожного користувача $ai \in A$. Від'ємні значення ri(bk) означають неприязнь, а позитивні значення виражають симпатію ai до продукту bk. Нижні значення $ri(bk)=\perp$ вказують на те, що a_i на має рейтингу bk.

Завдяки високій якості продукції та мінімальним вимогам до інформації, системи спільної фільтрації стали найбільш яскравими представниками рекомендаційних систем. Багато комерційних постачальників, наприклад, Amazon.com, використовують варіації методів колаборативної фільтрації, щоб пропонувати продукти своїм клієнтам. Крім простих байєсівських класифікаторів, хортинг та методики, засновані на правила асоціації, переважно два підходи набули широкого поширення, а саме, заснований на користувачах і на елементах на основі спільної фільтрації. Насправді, термін "спільна фільтрація" зазвичай використовується як синонім до "спільна фільтрація на основі користувачів", завдяки величезній популярності цієї техніки. Наступні два розділи приблизно описують алгоритмічні реалізації спільної фільтрації як на основі користувачів, так і на основі елементів.

4.3 Спільна фільтрація на основі користувачів

Проекти Ringo та GroupLens були одними з перших рекомендаційних систем, які застосовували методи, відомі як “спільна фільтрація на основі користувачів”. Представляючи функцію оцінки ai кожного користувача ri як вектор, вони спочатку обчислюють схожість $c(ai,aj)$ між усіма парами $(ai,aj) \in A \times A$. З цією метою використовуються загальні статистичні коефіцієнти кореляції, як правило, кореляція Пірсона і косинусна міра подібності, добре відома з пошуку інформації. Як випливає з назви, косинусна міра подібності кількісно визначає подібність двох векторів $v_i \rightarrow v_j$ $[-1,+1]$ за косинусом їхніх кутів:

$$sim(\vec{v}_i, \vec{v}_j) = \frac{\sum_{k=0}^{|B|} v_{i,k} \cdot v_{j,k}}{\left(\sum_{k=0}^{|B|} v_{i,k}^2 \cdot \sum_{k=0}^{|B|} v_{j,k}^2\right)^{\frac{1}{2}}}$$

Рис. 4.3 Визначення косинусної міри подібності

Кореляція Пірсона, отримана з моделі лінійної регресії, подібний до косинусної подібності, але вимірює ступінь лінійної залежності яка існує між двома змінними. Символи vi, vj позначають середні значення векторів vi, vj

$$sim(\vec{v}_i, \vec{v}_j) = \frac{\sum_{k=0}^{|B|} (v_{i,k} - \overline{v}_i) \cdot (v_{j,k} - \overline{v}_j)}{\left(\sum_{k=0}^{|B|} (v_{i,k} - \overline{v}_i)^2 \cdot \sum_{k=0}^{|B|} (v_{j,k} - \overline{v}_j)^2\right)^{\frac{1}{2}}}$$

Рис. 4.4 Кореляція Пірсона

Для обчислення подібності $c(ai,aj) \in A \times A$ за допомогою косинусної міри подібності або кореляції Пірсона, будуються околиці prox(ai) найбільш подібних сусідів top-M для кожного рівного $ai \in A$. Далі обчислюються передбачення для всіх добутків bk, які оцінили сусіди ai, але які ще невідомі ai,

тобто, більш формально, передбачення wi(bk) для $b_k \in \{b \in B \mid \exists\, a_j \in prox(a_i) : r_j(b) \neq \perp\}$:

$$\omega_i(b_k) = \overline{r}_i + \frac{\sum_{a_j \in prox(a_i)} (r_j(b_k) - \overline{r}_j) \cdot c(a_i, a_j)}{\sum_{a_j \in prox(a_i)} c(a_i, a_j)}$$

Рис 4.5

Таким чином, передбачення базуються на середньозважених значення відхилень від середніх значень сусідів ai. Для топ-N рекомендацій обчислюється список $Pwi : \{1,2,\ldots,N\} \rightarrow B$ на основі передбачення wi.

Зауважте, що функція Pwi є ін'єктивною і відображає рейтинг рекомендацій у порядку спадання, першими даючи найвищі прогнози. Щоб зробити кращі прогнози, різні дослідники запропонували кілька модифікацій основного алгоритму спільної фільтрації на основі користувачів. Нижче наведено найвідоміші з них, але, безумовно, їх є набагато більше.

Інверсна частота користувача. У програмах пошуку інформації, заснованих на моделі векторного простору, частоти слів зазвичай змінюються за фактором, відомим як "інверсна частота документа". Ідея полягає в тому, щоб зменшити вплив слів, що часто зустрічаються і збільшити вагу для незвичайних термінів під час обчислення подібності між векторами документа. Зворотна частота користувачів, вперше згадана Breese приймає це поняття і винагороджує співголосування за менш поширені товари набагато більше, ніж співголосування за дуже популярні продукти.

Зважування значимості. Обчислення кореляцій користувача та користувача c(ai, aj) враховує лише продукти, які обидва користувачі оцінили, тобто $b_k \in (\{b \mid r_i(b) \neq \perp\} \cap \{b \mid r_j(b) \neq \perp\})$.Таким чином, навіть якщо ai та aj мають кооперацію лише з один добутком bk, вони матимуть максимальну кореляцію, якщо виконується $r_i(b_k) = r_j(b_k)$. Очевидно, що такі кореляції, засновані лише на кількох точках даних, є не дуже надійними. Герлокер запропонував штрафувати кореляції користувачів на основі менш ніж 50

загальних оцінок, застосовуючи вагу значимості $\frac{s}{50}$, де s позначає кількість предметів спільної оцінки. Голосування за замовчуванням є ще одним підходом до вирішення тієї ж проблеми.

Корпус посилення. Хоча обидві попередні модифікації стосуються подібних процесів обчислення, посилення випадку розглядає крок прогнозування рейтингу, формалізованого у рівнянні 2.5. Підкреслюються коефіцієнти кореляції $c(ai, aj)$, близькі до одиниці, а низькі коефіцієнти кореляції караються:

$$c'(a_i, a_j) = \begin{cases} c(a_i, a_j)^{\rho}, & c(a_i, a_j) \geq 0 \\ -(-c(a_i, a_j))^{\rho} \end{cases}$$

Рис 4.6 Коефіцієнти кореляції

Отже, дуже схожі користувачі мають набагато більший вплив на прогнозовані рейтинги, ніж раніше. Зазвичай ρ приймає значення близько 2,5.

Деякі дослідники дещо розширили концепцію спільної фільтрації на основі користувачів і додали фільтр-ботів як додаткових користувачів, які можуть бути обрані як сусіди для “справжніх” користувачів. Filterbots - це автоматизовані програми, які поводяться певним, заздалегідь визначеним чином. Наприклад, у контексті рекомендації новин, деякі фільтр боти оцінювали татті Usenet на основі частки орфографічних помилок, тоді як інші зосереджувалися на довжині тексту тощо. Sarwar показа, що фільтр-боти можуть підвищити точність рекомендацій при роботі в малонаселених системах спільної фільтрації.

4.4 Спільна фільтрація на основі елементів

Спільна фільтрація на основі елементів набирає обертів протягом останніх п’яти років завдяки сприятливим характеристикам складності обчислень і здатності відокремити процес моделювання обчислень від реального прогнозування. Особливо для випадків коли $|A| > |B|$, було показано, що обчислювальна продуктивність спільної фільтрації на основі елементів перевищує спільну фільтрацію на основі користувачів. Його успіх також поширюється на багато комерційних рекомендаційних систем, таких як Amazon.com.

Як і спільна фільтрація на основі користувачів, створення рекомендацій базується на оцінках ri(bk), які користувачі $ai \in A$ надають для продуктів $bk \in B$. Однак, на відміну від спільної фільтрації на основі користувачів, значення подібності с обчислюються для елементів, а не для користувачів, отже $bk \in B$. Грубо кажучи, два елементи bk, be подібні, тобто мають велике значення c(bk, be), якщо користувачі, які оцінюють один з них, мають тенденцію оцінювати інший, і якщо користувачі схильні призначати їм ідентичні або подібні оцінки. Фактично, обчислення подібності на основі елемента прирівнюється до випадку на основі користувача при повороті матриці продукт-користувач на 90 градусів. Даля для кожного bk визначаються околиці $prox(bk) \subseteq B$ найдрібніших елементів топ-M wi(bk) обчислюються наступним чином:

$$w_i(b_k) = \frac{\sum_{b_e \in B'_k} (c(b_k, b_e) \cdot r_i(b_e))}{\sum_{b_e \in B'_k} |c(b_k, b_e)|}$$

Рис 4.7 Визначення околиць найподібніших елементів

Інтуїтивно, цей підхід намагається імітувати реальну поведінку користувача, коли користувач аі оцінює цінність невідомого продукту bk,

порівнюючи останній із відомими, подібними елементами та враховуючи, наскільки вони цінуються.

Остаточне обчислення списку Pwi з перших N рекомендацій слідує процесу спільної фільтрації на основі користувача, упорядковуючи рекомендації відповідно до wi в порядку спадання.

5 СПЕЦІАЛЬНА ЧАСТИНА

Це дослідження має на меті вирішити три основні цілі: побудова бази даних графів з даних виробів. досліджувати кластеризацію в базах даних виробів і досліджувати, наскільки добре ці кластери можна використовувати для рекомендації виріб. Аналіз був розбитий на три аспекти: дослідити, наскільки добре були побудовані графіки, дослідити, наскільки добре працювали алгоритми кластеризації, дослідити, як кластери працювали як метод рекомендації. Оскільки було досліджено дві структури графа, обидві були проаналізовані окремо. На одному позначеному графу було досліджено, наскільки добре працюють усі різні методи кластеризації, а на графі з кількома мітками було досліджено, як зміна включених вузлів вплинула на результати застосовних алгоритмів.

5.1 Одновузловий граф

Першою частиною аналізу було дослідження того, як будувався графік. Це включало перегляд кількості вузлів і зв’язків. Початковою ознакою успішного графіка є наявність вузла, створеного для кожного виробу, що містить необхідні властивості. Інша річ, яку спочатку слід розглянути, це те, чи вдало з’єднуються стосунки та чи є значуща кількість зв’язків. Наприклад, не було б сенсу, щоб вузол з’єднувався з кожним іншим вузлом для будь-яких властивостей. Також важливо перевірити, чи існує розумна кількість зв’язків на вузол.

Граф з одним вузлом був побудований шляхом спочатку встановлення всіх вузлів в одному запиті, а потім запиту на створення кожного зв’язку. Кожне відношення було засноване на перетині властивостей вузла. Вони додають по одному запиту.

У комп’ютері не вистачило оперативної пам’яті коли порівнювались країни, хоча для комп’ютера вона була максимальною – 7 ГБ. Це тому, що майже все виробляється в США. При спробі побудувати рекомендаційні системи втрата інформації про виріб може спричинити негативний ефект.

Однак у цьому випадку, включаючи країну походження, для рекомендаційної системи не є добре, оскільки такий загальний результат не може запропонувати відмінності між виробами. Також побудова спільного зв’язку між кожним вузлом може викликати плутанину в алгоритмах графів.

Кількість створених матеріальних стосунків становить 34,4567, що більше в 1,37 разів ніж очікувалося. Ця різниця є розумною на основі зроблених припущень. Більше значення також має сенс, оскільки більш популярні вироби мають менший вибір виробника. Зазвичай можна очікувати, що на один виріб буде від однієї і більше моделей. Друга дослідницька робота показала що очікувана кількість спільних відносин матеріалу на вузол буде від двох до шести. Досягнуте тут значення 3,3108 добре вписується в цей прогноз. Визначення призначення залежить від самого підприємства, тому важко порівняти ці цифри зі статистикою виробів. Однак вибране підприємство пропонує вибір лише з двадцяти матеріалів. Ці результати свідчать про те, що 45% виробів будуть мають однакові матеріали. Якщо порівняти кількість виробів у даних, 4803, з кількістю матеріалів, то зрозуміло, що буде велике перехрестя матеріалів.

Ключові слова підприємства охоплюють ряд властивостей виробів. Таким чином, для кожного вузла має бути велика кількість спільних ключових властивостей.

5.2 Основне зважування

При побудові системи з ваговими коефіцієнтами кожне співвідношення додавалося окремо. Така ж проблема з процесорною потужністю виникла при розгляді країни-виробника. Однак у цьому випадку вплинуло й кількість типів. Ця додаткова втрата інформації не є ідеальною під час спроби представити дані запису. Однак це принесло додаткову перевагу вагових показників графіка. Втрата конкретних деталей про виріб призводить до кращого уявлення про те, наскільки міцно пов’язані вузли.

Отримана кількість зв’язків була такою ж, як і на попередньому графіку. Це було очікувано, оскільки єдиною бажаною зміною були додані ваги. Це

показує, що доданий процес не вплинув на зміну структури графу. Однак час роботи цього зайняв у середньому в 11,6 разів більше часу. Зробивши його набагато більш складнішим у обчислювальних аспектах. Цей метод був успішним у створенні бази даних графів, що з’єднує вузли виробів із зваженими зв’язками, визначеними спільними властивостями запису.

Середня вага матеріалу становила 0,2048, що означає, що в середньому на перехресті був лише один матеріал, оскільки кількість матеріалів для кожного виробу була встановлена на п’ять. Кілька сильніших результатів допомагають створити міцніші стосунки, які можна відрізнити від основної кількості виробників.

Можна визначити що середня вага відносин виробників дорівнює 0,972 з дуже малим діапазоном. Це пов’язано з тим, що на виріб є лише виробників, а це означає, що будь-який перетин, швидше за все, буде таким же, як об’єднання властивостей, що дасть співвідношення одиниці.

Спільні ключові слова мають дуже мале середнє значення 0,165, хоча є деякі дуже сильні відхилення та більший розподіл, ніж інші. Компанія демонструє набагато більш рівномірний розподіл із середнім значенням 0,416.

Проблема з показаним контрастом у вагових показниках полягає в тому, що майже всі відносини з виробником переважають зв’язки з ключовими словами, що потенційно робить дані про ключові слова зайвими. Однак, оскільки кількість зв’язків спільного виробника на вузол набагато нижча, ніж кількість спільних зв’язків із ключовими словами на вузол, це означає, що зв’язки спільного виробника є більш корисним ідентифікатором зв’язків під час кластеризації.

5.3 Фіксовані зважування

При створенні графіка на основі частоти документа, властивості частоти були успішно створені в бажаному вигляді, як описано. Однак під час спроби налагодити стосунки виникли проблеми. Оскільки процес заснований на створенні таблиці, яка містить вузли для з'єднання та їх перетин. Через кількість можливих вузлів і з'єднань виготовлення цієї таблиці було занадто

вимогливим з точки зору обчислень. Хоча були спроби з усіма дослідженими властивостями, жоден з спроб не зміг створити таблицю без збоїв.

Щоб дослідити доказ концепції, методику випробували з графом з двадцяти виробів. Процес був таким же, але копія csv, що містить лише перші двадцять записів, була прочитана в Neo4j. Була створена невелика версія потрібного графу. Хоча на жаль, граф не може бути досліджений далі, підтвердження концепції показує його здатність використовувати частоту документа як вагові коефіцієнти для відносин.

5.3 Поєднання відносин

Цей набір містить 847 246 властивостей, створив 847 246 зв’язків і був завершений через 5 439 016 мс. Код успішно встановив одну властивість для кожного нового зв’язку. Це 176,4 відносини на вузол. в 2,18 рази більше середніх відносин на вузол від основного результату зважування. З огляду на те, що середні значення скошені екстремальними значеннями, це говорить про те, що всі відносини були успішно об’єднані. З цього зразка видно, що існує багато зв’язків на один вузол, однак є лише два зв’язки між будь-якими двома вузлами.

Можна помітити, що і середнє значення, і інтерквартильний діапазон дуже низькі. Однак існують дуже великі аномальні значення. Це може свідчити про ефективний потенціал кластеризації, оскільки кілька неймовірно сильних ваг встановлять чіткі зв’язки між багатьма слабкими зв’язками.

5.4 Граф з кількома вузлами

Аналіз розроблених графів із кількома мітками був розділений на дві частини, щоб відповідати першим двом основним цілям. Першою частиною було дослідження побудови графів, а потім застосування алгоритмів на основі спільноти.

5.5 Дослідження кластеризації

Оскільки структура цього графіка не дозволяла трикутникам формувати

алгоритми на основі трикутників, кількість трикутників, коефіцієнт зв’язності, компонент зв’язку та компонент сильної зв’язку були непридатними. Таким чином, кластеризаційний аналіз проводиться з використанням поширення мітки та модульності Lovain. Цей графік також не мав ваг у своїх відносинах, тому аспект алгоритмів кластеризації також не міг бути досліджений.

Кількість виробів у кожному кластері після зйомки режиму можна побачити на рисунку 5.7.

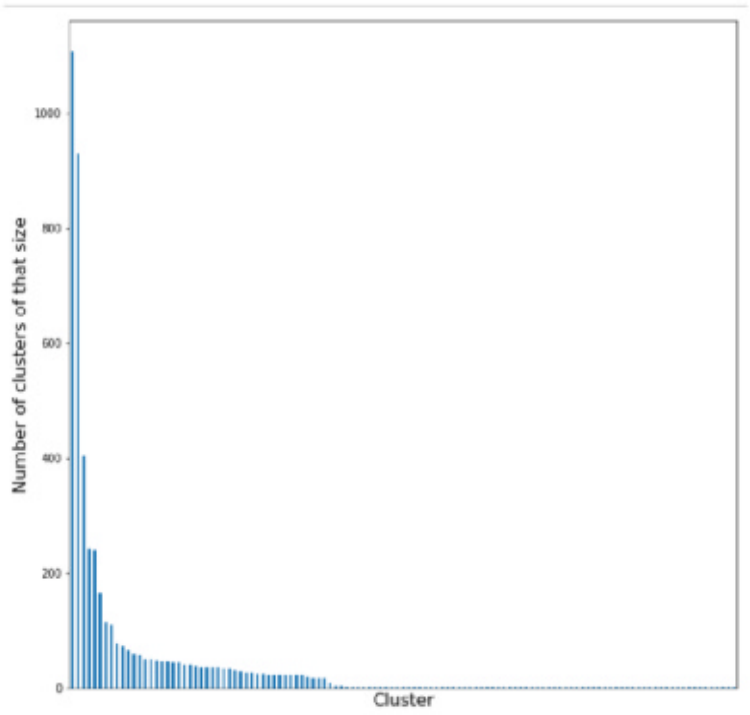

Рис. 5.7. Гістограма, що показує кількість виробів на кластер за допомогою модульності Лувена.

Цей алгоритм створив 678 кластерів із середнім показником 40,36. Як видно на цьому зображенні, у вузлах домінують в основному кілька стовпців, як і з результатами кластеризації однієї мітки. Як видно на цьому зображенні, у вузлах домінують в основному кілька стовпців, як і з результатами кластеризації однієї мітки. Найбільший кластер містив 1106 вироби. Це означає, що в одному кластері існує 0,23 вироби. Було 58 кластерів, які містили лише один виріб. При порівнянні цих результатів із викладеними показниками

кластеризації. Метод Лувена погано працює з цією структурою графіка як рекомендаційною системою. Кластери не утворюють рівномірного розподілу кластерів, що містять усі вироби.

Є занадто багато виробів, які не можна рекомендувати через те, що вони існують у власному кластері. Також важко дати значущі рекомендації, коли чверть усіх виробів вписується в один кластер.

Той самий метод був застосований до графу з поширенням мітки. Результати можна побачити нижче на рисунку 3.8

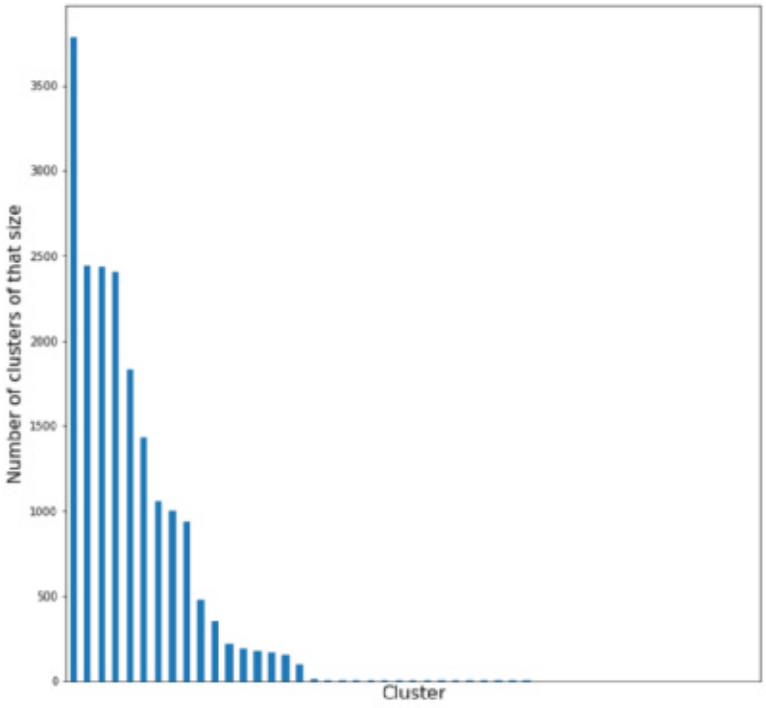

Рис. 5.14. Гістограма, що показує кількість виробів на кластер за допомогою поширення міток.

Найдивовижніше, якщо дивитися на цю стовпчасту діаграму, це те, що розмір кластерів в сумі набагато більше, ніж кількість виробів. Це очікувано, оскільки кластерів було набагато більше, ніж у виріб. Варто провести подальше дослідження, щоб визначити, чи розподіляються вироби рівномірно. Метою цього експерименту було дослідження окремо згрупованих виробів. Хоча, можливо, варто додатково дослідити цю форму кластеризації як метод рекомендації виробу. Навіть враховуючи це, розподіл розміру кластерів

занадто великий, щоб побудувати успішну систему рекомендацій, як було зазначено раніше.

5.6 Порівняння графів

Якщо порівняти кластеризацію двох графів, стає ясно, що жоден із використаних підходів не був успішним у створенні окремих кластерів. Однак найбільш близькою до ефективної виявилася модуляція Ловейна на графіку Multi-Label. Хоча неможливо було повністю дослідити, як зважена частота документа порівнялася б із показниками кластеризації, оскільки не було достатньо обчислювальної потужності для повного дослідження.

6 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

6.1 ОХОРОНА ПРАЦІ

6.1.1 Аналіз потенційних шкідливих впливів на працівників, що працюють з ЕОМ

Аналіз виробничого травматизму показує, що кількість травм, які спричинені дією електричного струму є незначною і складає близько 1 %, однак із загальної кількості смертельних нещасних випадків частка електротравм вже складає 20-40 % і займає одне з перших місць. Найбільша кількість випадків електротравматизму, в тому числі із смертельними наслідками, стається при експлуатації електроустановок напругою до 1000 В, що пов'язано з їх поширенням і відносною доступністю практично для кожного, хто працює на виробництві.

Основними причинами електротравматизму на виробництві є: випадкове доторкання до неізольованих струмопровідних частин електроустаткування; використання несправних ручних електроінструментів; робота без надійних захисних засобів та запобіжних пристосувань; доторкання до незаземлених корпусів електроустаткування, що опинилися під напругою внаслідок пошкодження ізоляції; недотримання правил улаштування, технічної експлуатації та правил техніки безпеки при експлуатації електроустановок, ЕОМ та інших електричних пристроїв.

Електроустаткування, з яким доводиться мати справу практично всім працівникам на виробництві, становить значну потенційну небезпеку ще й тому, що органи чуття людини не здатні на відстані виявляти наявність електричної напруги. В зв'язку з цим захисна реакція організму проявляється лише після того, як людина потрапила під дію електричної напруги. Проходячи через організм людини електричний струм справляє на нього термічну, електролітичну, механічну та біологічну дію.

В таблиці 6.1 наведено порогові значення змінного та постійного струму, що дуже важливо знати працівникам, які безпосередньо взаємодіють з електронною апаратурою чи мають до неї доступ.

Таблиця 6.1 - Порогові значення змінного та постійного струму

Вид струму	Пороговий відчутний	Пороговий невідпускаючий струм, мА	Пороговий фібриляційний струм, мА
Змінний струм частотою 50 *Гц*	0,5-1,5	6-10	80-100
Постійний струм	5,7-7,0	50-80	300

Шкіливий вплив на організм працівників, що працюють з ЕОМ також чинять різного роду електромагнітні випромінювання. Їх продукують монітори з електронно-променевою трубкою і в значно меншій мірі самі схеми пристрою. Наприклад в таблиці 6.2 подані види випромінювання електронно-променевих трубок і відповідні нормовані значення.

Таблиця 6.2 - Види випромінювання електронно-променевих трубок

Види випромінювань	**Діапазон**	**Фактичні (середні) дані замірів**	**Нормовані значення**
Рентгенівське	понад 1,2 КеВ	9-10-12 мкр/г	75,0 мкр/г
Ультрафіолетове випромінювання	220-280 нм 280-320 нм	0 мкр/г 0-0,02 мкр/г	0,01 Вт/м 0,01 Вт/м
Видимий діапазон	320-400 нм 400-700 нм	0,1-2,0 мкр/г 2,5-4,0 мкр/г	10 Вт/м
ІЧ-випромінювання	700 нм-1 мм	0,05-4,0 мкр/г	100 Вт/м
Електростатичне поле	0 Гц	15 кВ/м	20-60 кВ/м
Електричний струм	50 Гц	U=220В I=2А	U=220В I=0,1 А

Електромагнітне випромінювання може викликати біологічні та функціональні несприятливі ефекти в організмі людини. Функціональні ефекти проявляються у передчасній втомлюваності, частих болях голови, погіршенні сну, порушеннях центральної нервової та серцево-судинної систем. При систематичному опроміненні ЕМП спостерігаються зміни кров'яного тиску, сповільнення пульсу, нервово-психічні захворювання, деякі трофічні явища

(випадання волосся, ламкість нігтів та ін.). Сучасні дослідження вказують на те, що радіочастотне випромінювання, впливаючи на центральну нервову систему, є вагомим стрес-чинником.

Біологічні несприятливі ефекти впливу електромагнітного випромінювання проявляються у тепловій та нетепловій дії. Нині достатньо вивченою можна вважати лише теплову дію ЕМП, яка призводить до підвищення температури тіла та місцевого вибіркового нагрівання органів та тканин організму внаслідок переходу електромагнітної енергії у теплову. Таке нагрівання особливо небезпечне для органів із слабкою терморегуляцією (головний мозок, око, нирки, шлунок, кишківник, сім'яники). Наприклад, випромінювання сантиметрового діапазону призводять до появи катаракти, тобто до поступової втрати зору.

Механізм та особливості нетеплової дії ЕМП радіочастотного діапазону ще до кінця не з'ясовані. Частково таку дію пояснюють специфічним впливом радіочастотного випромінювання на деякі біофізичні явища: біоелектричну активність, що може призвести до порушення усталеного протікання хімічних та ферментативних реакцій, вібрацію субмікроскопічних структур, енергетичне збудження (часто резонансне) на молекулярному рівні, особливо на конкретних частотах.

Не слід забувати і про шум. Особливу увагу привертають ЕОМ старіших модифікацій, де спостерігається доволі високий показник шумності. Це пов’язано з використанням великої кількості охолоджуючих вентиляторів (кулерів), а також і з недосконалістю самих електричних вузлів ЕОМ.

Так за даними медиків дія шуму може спричинити нервові, серцево-судинні захворювання, виразкову хворобу, порушення обмінних процесів та функціонування органів слуху тощо.

Важливим фактором є освітленість. Якщо вона недостатня чи навпаки спостерігається надлишок – то це може призвести до великих проблем із зором.

Адже відомо, що майже, 90% всієї інформації про довкілля людина одержує через органи зору. Під час здійснення будь-якої трудової діяльності

втомлюваність очей, в основному, залежить від напруженості процесів, що супроводжують зорове сприйняття. При поганому освітленні людина швидко втомлюється, працює менш продуктивно, зростає потенційна небезпека помилкових дій і нещасних випадків. Основними причинами поганої освітленості: погане планування розміщення вікон, погане планування штучного освітлення, використання неякісних моніторів, що зумовлюється не виконанням норм.

6.1.2 Основні вимоги з охорони праці до користувачів ЕОМ та їх робочого місця

При експлуатації ПК необхідно пам'ятати, що первинні мережі електроспоживання під час роботи знаходяться під напругою, яка є небезпечною для життя людини, тому необхідно користуватися справними розетками, відгалужувальними та з'єднувальними коробками, вимикачами та іншими електроприладами. До роботи з ПК допускаються працівники, з якими проведений вступний інструктаж та первинний інструктаж (на робочому місці) з питань охорони праці, техніки безпеки, пожежної безпеки та зроблений запис про їх проведення у спеціальному журналі інструктажів. Працівники при роботі з ПК повинні дотримуватися вимог техніки безпеки, пожежної безпеки. При виявленні в обладнанні ПК ознак несправності (іскріння, пробоїв, підвищення температури, запаху гару, ознак горіння) необхідно негайно припинити роботи, відключити усе обладнання від електромережі і терміново повідомити про це відповідних посадових осіб, спеціалістів. Потрібно знати місця розташування первинних засобів пожежегасіння, план евакуації працівників, матеріальних цінностей з приміщення в разі виникнення пожежі.

Перед початком роботи на ПК користувач повинен:

- пересвідчитися у цілості корпусів і блоків (обладнання) ПК;
- перевірити наявність заземлення, справність і цілість кабелів живлення, місця їх підключення.

Забороняється вмикати ПК та починати роботу при виявлених несправностях.

Під час роботи, пересвідчившись у справності обладнання, увімкнути електроживлення ПК, розпочати роботу, дотримуючись умов інструкції з її експлуатації.

Великий вплив на умови праці здійснює приміщення, в якому безпосередньо працюють люди з ЕОМ. До таких приміщень є ряд вимог:

- стіни приміщень для роботи з ПК мають бути пофарбовані чи обклеєні шпалерами пастельних кольорів з коефіцієнтом відбиття 40 - 60 %. У випадках, коли такі приміщення зорієнтовані на південь, вікна повинні обладнуватися сонцезахисними пристроями (жалюзі, штори і т. п.);
- для освітлення приміщень з ПК необхідно використовувати люмінесцентні світильники. Освітленість робочих місць у горизонтальній площині на висоті 0,8 м від підлоги повинна бути не менше 400 лк. Вертикальна освітленість у площині екрану не більше 300 лк;
- у приміщеннях для роботи з ПК необхідно проводити щоденне вологе прибирання та регулярне провітрювання протягом робочого дня.

Також висуваються окремі вимоги до робочого місця:

- робочі місця для працюючих з дисплеями необхідно розташовувати таким чином, щоб до поля зору працюючого не потрапляли вікна та освітлювальні прилади. Відео термінали повинні встановлюватися під кутом 90 - 105 градусів до вікон та на відстані, не меншій 2,5 - 3 м від стіни з вікнами;
- до поля зору працюючого з дисплеєм не повинні потрапляти поверхні, які мають властивість віддзеркалювання. Покриття столів повинне бути матовим з коефіцієнтом 0,25 - 0,4;
- відстань між робочими місцями з ПК повинна бути не меншою 1,5 м у ряду та не меншою 1 м між рядами. ПК повинні розміщуватися не ближче 1 м від джерела тепла;

- відстань від очей користувача до екрану повинна становити 500 - 700 мм, кут зору - 10 - 20°, але не більше 40°, кут між верхнім краєм відео терміналу та рівнем очей користувача повинен бути меншим 10°. Найбільш вигідне є розташування екрану перпендикулярно до лінії зору користувача;
- монітор обладнується захисною плівкою, що розсіює шкідливе електромагнітне випромінювання.
- оптимальні розміри робочої поверхні стільниці 1600×900 мм. Під стільницею робочого столу повинно бути вільний простір для ніг із розмірами по висоті не менше 600мм, по ширині 500мм, по глибині 650мм;
- всі ЕОМ повинні бути заземлені.

6.1.3 Безпечна експлуатація електроустановок та пожежна безпека

Робота щодо забезпечення безпечної експлуатації електроустановок здійснюється згідно з обов'язковими, для всіх споживачів електроенергії, незалежно від їх відомчої приналежності, правилами технічної експлуатації електроустановок споживачів та правилами техніки безпеки при експлуатації електроустановок споживачів. Обслуговування діючих електроустановок, проведення в них оперативних переключень, організація та виконання ремонтних, монтажних, налагоджувальних робіт і випробувань здійснюються спеціально підготовленим електротехнічним персоналом.

Безпечна експлуатація електроустановок забезпечується: конструкцією електроустановок; технічними способами та засобами захисту; організаційними та технічними заходами.

Конструкція електроустановок повинна відповідати умовам їх експлуатації та забезпечувати захист персоналу від можливого доторкання до рухомих та струмовідних частин, а устаткування - від потрапляння всередину сторонніх предметів та води.

За способом захисту людини від ураження електричним струмом встановлено п'ять класів електротехнічних виробів: 0, 01,1, II, III. До класу 0 належать вироби, які мають робочу ізоляцію і у яких відсутні елементи для заземлення. До класу 01 належать вироби, які мають робочу ізоляцію, елемент для заземлення та провід без заземлювальної жили для приєднання до джерела живлення. До класу I належать вироби, які мають робочу ізоляцію та елемент для заземлення. Якщо виріб класу I має кабель до джерела живлення, то цей кабель повинен мати заземлювальну жилу та штепсельну вилку зі заземлювальним контактом. Цей контакт є дещо довшим за робочі контакти вилки для того, щоб забезпечувати випереджальне замикання заземлювального контакту під час увімкнення та більш запізніле розмикання його під час вимикання. До класу II належать вироби, які мають подвійну чи посилену ізоляцію і не мають елементів для заземлення. До класу III належать вироби, які не мають внутрішніх та зовнішніх електричних кіл з напругою понад 42 В.

Технічні способи та засоби захисту (ТСЗЗ) поділяють на:

1) ТСЗЗ при нормальних режимах роботи електроустановок (ізоляція струмовідних частин, забезпечення недоступності неізольованих струмовідних частин, попереджувальні сигналізація, знаки та написи, застосування малих напруг, захисне розділення електромереж, вирівнювання потенціалів);

2) ТСЗЗ при переході напруги на металеві нормально неструмовідні частини електроустановок (захисні заземлення, занулення, вимикання);

3) електрозахисні засоби та запобіжні пристосування.

Забезпечення пожежної безпеки — це один із важливих напрямків щодо охорони життя та здоров'я людей.

Основною причиною пожеж у приміщеннях, де використовуються ЕОМ є в основному короткі замикання, які виникають внаслідок неправильного монтажу або екс-плуатації електроустановок, старіння або пошкодження ізоляції. Струм короткого замикання залежить від потужності джерела струму, відстані від джерела струму до місця замикання та виду замикання. Великі струми замикання викликають іскріння та нагрівання струмопровідних частин

до високої температури, що може викликати займання ізоляції провідників та горючих будівельних конструкцій, які знаходяться поряд.

Ще однією причиною є струмові перевантаження, що виникають при ввімкненні до мережі додаткових споживачів струму або при зниженні напруги в мережі. Тривале перевантаження призводить до нагрівання провідників, що може викликати займання ізоляції.

Продуктами згорання стають багато речовин, високі концентрації яких можуть серйозно впливати на організм людини (таблиця 6.3)

Таблиця 6.3 – Степені небезпечності концентрацій продуктів згорання

Речовини	Концентрація					
	смертельна за умови вдихання протягом 5 - 10 хв.		небезпечна (отруйна) за умови вдихання протягом 0,5-1,0 год.		переносима за умови вдихання протягом 0,5-1,0 год.	
	%	г/м³	%	г/м³	%	г/м³
Оксид азоту	0,05	1,0	0,01	0,2	0,005	0,1
Оксид вуглецю	0,5	6,0	0,2	2,4	0,1	1,2
Вуглекислий газ	9,0	162	5,0	90	3,0	54
Сірчаний газ	0,3	8,0	0,04	ІД	0,01	0,3
Сірководень	0,08	1,1	0,04	0,6	0,02	0,3
Сірковуглець	0,2	6,0	0,1	3,0	0,05	1,5
Хлористий вуглець	0,3	4,5	0,1	1,5	0,01	0,15
Синильна кислота	0,02	0,2	0,01	0,1	0,005	0,05

Дим являє собою велику кількість видимих найдрібніших твердих та (або) рідинних часточок незгорівших речовин, що знаходяться в газах у завислому стані. Він викликає інтенсивне подразнення органів дихання та слизових оболонок (сильний кашель, сльозотечу тощо). Крім того, у задимлених приміщеннях внаслідок погіршення видимості сповільнюється евакуація людей, а часом провести її зовсім неможливо.

При пожежі електроустаткування в приміщеннях використовуються вуглекислотні вогнегасники типу ОУ-2, ОУ-5, ОУ-8 ємністю 2,5 - 8 літрів, які призначені для гасіння пожеж всіх видів. Також невелику ділянку пожежі

можна локалізувати методом пониження доступу кисню в осередок вогню накинувши на нього азбестове полотно або грубу шерстяну тканину. Перелік засобів гасіння або локалізації вогню наводяться в таблиці 6.4.

Таблиця 6.4 – Засоби гасіння вогню

Назва приміщення	Площа, яка захищається, м2	Типи первинних засобів пожежогасіння	Кількість, шт.
Комп'ютерний зал	100	Вуглекислотні вогнегасники типу ОУ-8	2
		Азбестове полотно (або кошма) 1×1, 2×1 або 2×2 м	4

Пожежний захист і вибухозахист забезпечуються правильним вибором ступеня вогнестійкості окремих елементів і конструкцій; обмеженням розповсюдження вогню у випадку виникнення пожежі; впровадженням систем активного поглинання вибуху; застосування систем протидимового захисту; забезпеченням безпечної евакуації людей; застосуванням засобів пожежної сигналізації, оповіщення і пожежогасіння; організацією пожежної охорони об'єкта.

Обов'язковим є інструктаж персоналу з пожежної безпеки .

6.1.4 Характеристика та розрахунок захисного заземлення електроустановок

Захисне заземлення – допоміжне електричне з'єднання з землею чи її еквівалентом металевих неструмоведучих частин, що можуть виявитися під напругою.

Мета захисного заземлення – знизити напругу дотику між корпусом електроустановки і землею до 42В, і менше що там виникає в результаті ушкодження чи пробою ізоляції струмоведучих частин.

Захисне заземлення варто відокремити від робітника і заземлення для захисту від розрядів статичної та атмосферної електрики.

Робоче заземлення – допоміжне з'єднання з землею нейтральних точок обмоток генераторів, силових і вимірювальних трансформаторів, дугогасних

апаратів та інших ланцюгів з метою забезпечення нормальної роботи електроустановок. Заземлення для захисту від розрядів статичної й атмосферної електрики здійснюється для відводу цих зарядів у землю [10].

Розглянемо принцип роботи захисного заземлення. На рисунку 6.1,а показано ситуація дотик людини до заземленого корпуса електроустановки, на якому з’явилася напруга, а на рисунку 6.1,б – її еквівалентна електрична схема.

Спочатку визначимо значення напруги дотику Uдот, що прикладається до людини при дотику її до заземленого корпуса, з одного боку, і до ніг, з іншого, а потім значення струму I, що протікає через людину в цьому ланцюзі.

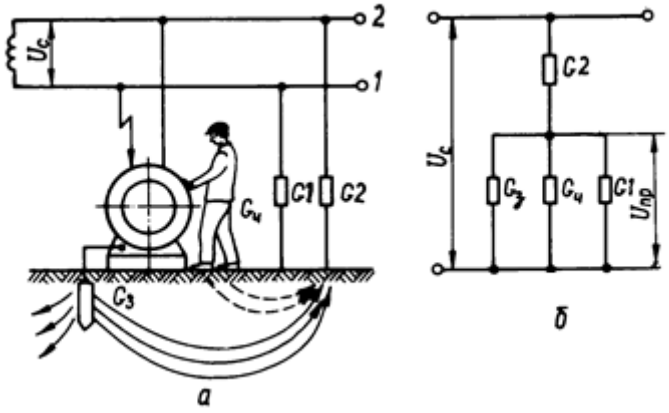

Рис. 6.1 – Еквівалентна електрична схема
а – схема дотику; б – еквівалентна електрична схема заземленої електроустановки

Для спрощення математичних перетворень оперувати будемо провідностями, а потім їх замінимо опорами.

Провідність заземлення Gз, провідність людини Gл і провідність ізоляції G1 проводу 1 щодо землі включені паралельно між собою і послідовно з провідністю G2 ізоляції проводу 2.

Сумарна провідність паралельного ланцюга складе:

$$G_{лан} = G_з + G_л + G_1 .$$

Провідність усього ланцюга

$$G=\frac{G_{лан}G_2}{G_{лан}+G_2}=\frac{(G_з+G_л+G_1)G_2}{G_з+G_л+G_1+G_2}$$

Напруга Uдот, що впливає на людину при дотику до корпусу електроустановки

$$\frac{U_{дот}}{U_м}=\frac{G}{G_{лан}}=\frac{(G_з+G_л+G_1)G_2}{(G_з+G_л+G_1+G_2)(G_л+G_1+G_з)}=\frac{G_2}{G_з+G_л+G_1+G_2}$$

де Uм – напруга мережі, В.

Тоді

$$U_{дот}=\frac{U_м G_2}{(G_з+G_л+G_1+G_2)}$$

Провідності Gл, G1, G2 набагато менше провідності заземлення Gз і ними як доданками в знаменнику можна знехтувати. Замінюючи провідності опорами і приймаючи r2 = rіз (rіз – опір ізоляції), одержимо

$$U_{дот}=\frac{U_м r_з}{r_{із}}=I_з r_з \qquad (6.1)$$

де Iз = Uм/rіз.

Аналіз виразу (6.1) дозволяє стверджувати, що найбільш доступним заходом щодо зниження напруги Uдот є зменшення опору заземлення rз, а збільшувати опір ізоляції економічно недоцільно.

Струм, що протікає через людину при дотику її до заземленого корпуса електроустановки [10]:

$$I_л=\frac{U_{дот}}{R_л}=\frac{U_м r_з}{R_л r_{із}} \qquad (6.2)$$

Виконуємо розрахунок захисного заземлення електроустановки.

Визначаємо кількість заземлювачів для заземлюючого пристрою підстанції. Зі сторони U1=11 кВ нейтраль ізольована, струм замикання на землю Iз=12А. На стороні U2=0,6 кВ нейтраль глухо заземлена. Опір природних заземлювачів Rпр=13Ом. Питомий опір ґрунту ρвим=0,9·104.

Заземлювачі вертикальні стрижневі діаметром 10мм довжиною 5м, коефіцієнт підвищення опору ґрунту ψ=2.

1. Визначаємо необхідний нормативний опір заземлюючого пристрою:

• для мережі 11 кВ,

$$R_з = \frac{U_з}{I_з} = \frac{125}{12} = 10,4 Ом$$

• для мережі 0,6 кВ, згідно ПУЕ $R_з \leq 4$ Ом .

Приймаємо менше значення опору Кз=4 Ом, оскільки заземлюючий пристрій спільний, як для мережі 11 кВ так і для 0,6 кВ.

2.Визначаємо опір штучних заземлювачів:

$$R_{шт} = \frac{R_з R_{пр}}{R_{пр} - R_з} = \frac{4 \cdot 13}{13 - 4} = 5,78$$ Ом.

В якості штучних заземлювачів приймаємо стрижневі електроди діаметром 10 мм і довжиною 5 м.

Визначаємо опір розтіканню струму одного заземлювача:

R0(рс)=0,00227·ρвим·ψ = 0,00227·0,9·104·2 = 40,86Ом.

3.Визначаємо кількість заземлювачів:

$$R_{шт} = \frac{R_з R_{пр}}{R_{пр} - R_з} = \frac{4 \cdot 13}{13 - 4} = 5,78$$ шт.

де Кb=0,56 – коефіцієнт використання електродів.

Приймаємо n=13 електродів.

4. Перевіряємо загальний опір заземлення:

$$R_{шт} = \frac{R_з R_{пр}}{R_{пр} - R_з} = \frac{4 \cdot 13}{13 - 4} = 5,78$$ Ом;

$$R_з = \frac{R_{пр} \cdot R_{шт}}{R_{пр} + R_{шт}} = \frac{13 \cdot 5{,}61}{13 + 5{,}61} = 3{,}92$$ Ом.

Значить кількість електродів вибрана вірно

6.2 БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

Заходи з безпеки в надзвичайних ситуаціях на об’єктах енергопостачання

Електропостачання міст і промислових підприємств здійснюється від власних джерел електростанцій або від потужних енергетичних систем.

Енергетичною системою називають групу електростанцій, зв'язаних електричними мережами між собою і з споживачами електроенергії. Енергетична система об'єднує різні по потужності і характеру електростанції (теплові, гідравлічні, атомні), працюючі на загальну мережу, що складається з ліній електропередач, трансформаторних підстанцій і інших споруд.

Енергосистема незалежно від віддалення електростанцій, що входять в неї, і протяжність мереж являє собою єдине ціле, пов'язане спільністю режиму роботи, безперервністю процесу виробництва і розподілу електроенергії. Енергосистеми дозволяють більш повно і економічно використати енергетичні ресурси і можливості електростанцій з урахуванням характеру роботи споживачів електроенергії.

Електричні мережі і споруди можна поділити на дві категорії:

- електростанції і споруди системного значення;
- електричні споруди і мережі загального користування.

До складу мереж і споруд систем енергопостачання входять великі електростанції (теплові, гідравлічні, атомні), лінії електропередач напругою 110 кВ і більше і зв'язані з ними мережні споруди. Електроенергія від таких станцій подається в енергосистему і великим споживачам.

Електричні споруди і мережі загального користування призначені для живлення електроенергією, що отримується від енергосистеми і власних електростанцій, міських споживачів, в тому числі і дрібних промислових підприємств. Вони складаються з трансформаторних підстанцій, розподільних пунктів, кабельних (рідше повітряних) ліній і інших споруд, за допомогою

яких електроенергія трансформується до напруги 10кВ-380/220В і доставляється споживачам.

У випадку ядерного вибуху в місті електрична станція і зв'язані з нею споруди отримають різні по характеру руйнування і пошкодження. Велика енергосистема, що базується на великій кількості злектростанстій віддалених одна від іншої на значні відстані, і що має систему автоматичних пристроїв, здатних вмить відключити будь-яке джерело енергії і відповідні потужності споживачів і тим самим зберегти працездатність системи, є досить надійною Можливість повного виходу з ладу такої енергосистеми навіть при застосуванні ядерної зброї по багатьох містах і енергоджерелам одночасно малоймовірна. Найвразливішими елементами енергосистеми є наземні споруди (станції, підстанції, розподільні пункти, трансформаторні станції і інш).

Аварійно-відновні роботи на системах електропостачання міст проводяться в осередку ураження в цілях:

- відключення окремих ліній і дільниць мережі електропостачання в місцях проведення рятувальних робіт для забезпечення безпеки людей і запобігання утворенню пожеж;
- подачі електроенергії в окремі райони і дільниці осередку ураження;
- забезпечення електроенергією особливо важливих споживачів у разі часткового пошкодження ліній електропередач і джерел електроживлення.

Відключення окремих дільниць мережі електропостачання є необхідним в місцях проведення рятувальних робіт, де пошкоджені мережі низької напруги живляться від високовольтних ліній, що збереглися. Відключення проводиться шляхом вимкнення рубильників, з допомогою раз'єднювачів або перерізанням проводів. При пошкодженні високовольтних ліній електропередач вони автоматично вимикаються на найближчих понижуючих трансформаторних підстанціях (масляні або повітряні вимикачі) або на розподільних пунктах.

Подача електроенергії в окремі райони або дільниці осередку ураження

може бути необхідна для самих різних цілей: освітлення території на об'єктах робіт; живлення електродвигунів різних машин і електрифікованого інструмента з використанням яких проводяться рятувальні роботи; забезпечення роботи тимчасово розгорнутих або тих, що збереглися, медичних установ і для багатьох інших цілей. Подавати електроенергію в цих випадках найбільш доцільно по електролініям, що збереглися, якщо об'єми відновних робіт невеликі, або по прокладуваних тимчасових кабельних мережах із живленням від сусідніх джерел (трансформаторних підстанцій, кабельних мереж, що збереглися і від інших місць підключення).

Вихід з ладу системи електропостачання міста (розпад системи) навіть в умовах мирного часу грозить серйозними наслідками.

Всі встановлені на енергоблоках ТЕС теплові захисти, блокування, сигналізація, контрольно-вимірювальні прилади, автоматичні регулятори і електроприводи запірної арматури повинні бути нормально введені в роботу. Дії оперативного і ремонтно-налагоджувального персоналу ЦТАВ повинні бути направлені на попередження і усунення неполадок у вищенаведених схемах, які можуть привести до їх відмови в роботі або помилковій роботі

а) Аварією з вини оперативного і ремонтного персоналу ЦТАВ вважається:

1. пошкодження основного обладнання внаслідок помилкової роботи, відмови в роботі, помилок при перемиканнях, випробовуваннях і ремонтах, що виконується персоналом ЦТАВ;

2. повне скидання навантаження усіма енергоблоками ТЕС внаслідок причин, вказаних в пункті "а.1", навіть при умові збереження їх власних потреб і незалежно від тривалості скидання.

б) Браком з вини оперативного і ремонтного персоналу ЦТАВ вважається:

1. відключення енергоблока внаслідок помилкової роботи захисту через вихід з ладу давачів, реле і їх кабельно-комутаційних схем, але без урахування

їх механічного пошкодження сторонніми предметами і особами, залиття водою, обтікання парою, порушення щільності первинних вентилів, тривалої роботи при підвищеній температурі з відома оперативного персоналу КТЦ;

2. відмова в роботі схеми теплових захистів, що привела до посилювання аварійного положення на енергоблоці по причинах, вказаних в пункті "б.1";

3. пошкодження основного і допоміжного обладнання внаслідок помилкової роботи або відмови в роботі схем теплових захистів, якщо це не вважається аварією;

4. затримка пуску блоку через неготовність до работи схем теплових захистів внаслідок відмови в роботі або виходу з ладу давачів, реле або комутаційно-кабельних зв'язків, але без урахування їх механічного пошкодження сторонніми предметами і особами, залиття водою, обтікання парою, порушення щільності первинних вентилів, тривалої роботи при підвищеній температурі, якщо затримка пуску блоку привела до порушення диспетчерського графіка і створення небалансу навантажень в енергосистемі.

в) Аварією з вини оперативного персоналу КТЦ вважається:

1. пошкодження основного обладнання внаслідок помилкової роботи захисту через вихід з ладу давачів, кабельних зв'язків захистів внаслідок їх механічного пошкодження при прибиранні, обтіканні парою, заливанні водою, порушення щільності первинних вентилів, тривалої роботи при підвищеній температурі, якщо про це був передчасно повідомлений оперативний персонал КТЦ і не прийняв заходів для їх усунення;

2. пошкодження основного обладнання внаслідок самовільного виведення накладок, ключів захисту або необгрунтованої команди на виведення захисту з роботи;

3. повне скидання навантаження всіма енергоблоками цеху внаслідок причин, вказаних в пунктах "в.1", "в.2" навіть при умові збереження їх власних потреб, незалежно від тривалості скидання.

г) Браком з вини оперативного персоналу КТЦ вважається:

1. відключення енергоблока внаслідок помилкової роботи захисту по причині, вказаній в пункті "в.1";

2. посилювання аварійного положення на блоці внаслідок самовільного виведення накладок, ключів захисту або необгрунтованої команди на виведення захисту з роботи;

3. необгрунтоване або помилкове дистанційне відключення блоку за допомогою кнопки зупинки;

4. відключення енергоблока захистом внаслідок порушення його режиму роботи через помилки в регулюванні параметрів, порушення виробничих інструкцій і режимних карт;

5. пошкодження основного і допоміжного обладнання внаслідок помилкової роботи захисту по причинах, вказаних в пунктах "в.1", "в.2", якщо це не кваліфікується як аварія;

6. затримка пуску блоку через неготовність до роботи схем теплових захистів внаслідок пошкодження їх схем сторонніми предметами, залиття водою, обтікання парою, порушення щільності первинних вентилів, тривалої роботи при підвищеній температурі, якщо про це був своєчасно повідомлений оперативний персонал КТЦ, при умові, що затримка пуску блоку привела до порушення диспетчерського графіка, створення небалансу навантажень в енергосистемі.

ВИСНОВКИ

Двома основними цілями цього дослідження були розробка бази даних графів для представлення даних виробів та застосування до неї алгоритмів кластеризації для пошуку кластерів, які потім можна було б використовувати для системи рекомендацій. Перша мета була успішною, оскільки було побудовано багато різновидів графіків. Друге завдання виявилося не таким успішним. Хоча було багато різних підходів, був знайдений метод кластеризації бази даних графів, щоб її можна було використовувати для системи рекомендацій. Після застосування кількох алгоритмів кластеризації жодного результату не вдалося задовольнити метрикам, які були встановлені, щоб вважати метод придатним для рекомендаційної системи.

Neo4j виявився ефективним інструментом для побудови графіків і застосування алгоритмів кластеризації. Був використаний для успішної розробки графічної бази даних з описом виробів. Більшість інформації з бази даних виробів успішно перенесено в базу даних графів. Проте велика частина описаних даних не підходила для графічної бази даних. Однією з проблем, яка виникла, була відсутність обчислювальної потужності, яка перешкоджала розробці більш складних ваг. Подальше дослідження різних методів зважування термінів частоти було б корисним, якщо є доступ до більш продуктивних комп’ютерів.

Розглядаючи побудову графіків, як базовий, невагомий одновузловий граф, так і графи з кількома мітками були успішно побудовані за допомогою Neo4j. Обидва ці графіки можна було запитувати за допомогою мови запитів Neo4j Cypher. Хоча графік з одним вузлом відображав більше інформації про кожен виріб у його вузлах і зв’язках.

Граф з одним вузлом можна було адаптувати для додавання ваги зв’язкам. Це неможливо було зробити за допомогою графу з кількома мітками. Проте були запропоновані потенційні розробки для подальшого вивчення цього. Через це немає жодних подальших розробок багатомітного графа. Однак

для графа окремих вузлів три графи, які були успішно побудовані в Neo4j, які містили всю інформацію про базу даних і один, який успішно створив вибірку. Виходячи з цих результатів, для розробки баз даних графів краще використовувати граф з однією міткою.

На відміну від графа з кількома мітками, на графі з однією міткою можна було виконати всі обговорювані алгоритми для всіх, крім терміну, через його розмір. Хоча запропоновані потенційні зміни дозволять це зробити. Однак це був результат як поширення мітки, так і модульності Лувена, який працював краще, ніж будь-який результат у графі єдиної мітки при використанні рівного розподілу як метрики. Хоча це ще не може бути по-справжньому підтверджено, оскільки кластери містять не лише вузли, які представляють вироби. Лише групування цих вузлів можна використовувати для дослідження рекомендацій виробу.

У цій роботі досліджується кластеризація графів виробів з кінцевою метою розробки системи, яку можна використовувати як систему рекомендацій та швидкого пошуку. Однак через те, що застосовувані методи кластеризації не дають результатів, які задовольняють встановленим показникам, не було можливості експериментально визначити, чи будуть ці рекомендаційні вироби в одному кластері працювати як функціональна рекомендаційна система. Це можна зробити шляхом аналізу другого набору даних, який містить оцінки результатів попередніх пошуків. Це було зроблено за допомогою даних, зібраних з попередніх рекомендацій, оцінених людиною позитивно, і визначення їх ідеального кластера. Дані містять ідентифікатор користувача, назву виробу, оцінку 1–5 із кроком 0,5, яка описує задоволення від рекомендації. Оцінка 5 означає максимальне задоволення, а оцінка 1 означає мінімальне задоволення. Позитивні рекомендації, визначені як оцінка 3+, можуть бути показником рекомендації, яка сподобалась. Кластер з найбільшою кількістю позитивних оцінок користувачів буде використовуватися як кластер рекомендацій. Для експериментального дослідження цього слід використовувати розділення тестової послідовності.

Хоча результати дослідження кластеризації для рекомендацій виробів були не переконливими, у цій області ще багато можливої роботи.

ПЕРЕЛІК ПОСИЛАНЬ

1. Al Hasan, M., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security.
2. Angles, R. and Gutierrez, C. (2008). Survey of graph database models. ACM Computing Surveys, 40
3. Assefi,M.,Liu,G.,Wittie,M.P.,andIzurieta,C.(2015).Anexperimental evaluation of apple siri and google speech recognition. Proccedings of the 2015 ISCA SEDE
4. Bar-Ilan, J. and Peleg, D. (1991). Approximation algorithms for select- ing network centers.
5. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008
6. Bobadilla,J.,Ortega,F.,Hernando,A.,andBernal,J.(2012).Acollabo- rative filtering approach to mitigate the new user cold start problem.
7. Brusilovsky, P. and Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems.
8. Buerli, M. and Obispo, C. (2012). The current state of graph databases.
9. Businessweek, B. (2013).
10. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2009)
11. Chung, H., Park, J., and Lee, S. (2017). Digital forensic approaches for amazon alexa ecosystem. Digital Investigation
12. Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks.
13. Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components.

14. Dial, R. B. (1969). Algorithm 360: Shortest-path forest with topological ordering [h]
15. Duda, R. O. and Hart, P. E. (2001). Dg stork pattern classification. John Wiely and Sons.
16. Fatemi, M. and Tokarchuk, L. (2013). A community based social recommender system for individuals & groups.
17. Fleischer, L. K., Hendrickson, B., and Pınar, A. (2000). On identifying strongly connected components in parallel
18. Follows, S. (2013). How many films in an average film career. 20.Fouss, F., Yen, L., Pirotte, A., and Saerens, M. (2006). An experimental investigation of graph kernels on a collaborative recommendation task.
19. Fredman, M. L. and Tarjan, R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms.
20. GroupLens (2018). Movielens 20m dataset.
21. Gubichev, A., Bedathur, S., Seufert, S., and Weikum, G. (2010). Fast and accurate estimation of shortest paths in large graphs. In Proceedings of the 19th ACM international conference on Information and knowledge management
22. Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Ham- mers, A. (2006). Automatic anatomical brain mri segmentation combining label propagation and decision fusion. NeuroImage
23. Holzschuher, F. and Peinl, R. (2013). Performance of graph query languages: comparison of cypher, gremlin and native access in neo4j. In Proceedings of the Joint EDBT/ICDT 2013 Workshops
24. Huang, Z., Chung, W., Ong, T.-H., and Chen, H. (2002). A graph-based rec- ommender system for digital library
25. Kannan, R., Vempala, S., and Vetta, A. (2004). On clusterings: Good, bad and spectral.
26. Karlgren, J. (1994). Newsgroup clustering based on user behavior-a recommen- dation algebra.

27. Kemper, C. (2015). Beginning Neo4j. Springer.
28. Kleinberg, J. M. (2002). Small-world phenomena and the dynamics of information.
29. Lakiotaki, K., Matsatsinis, N. F., and Tsoukias, A. (2011). Multicriteria user modeling in recommender systems.
30. Lalwani, D., Somayajulu, D. V., and Krishna, P. R. (2015). A community driven social recommendation system.
31. Lam, X. N., Vu, T., Le, T. D., and Duong, A. D. (2008). Addressing cold-start problem in recommendation systems.
32. Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks.
33. Schaeffer, S. E. (2007). Graph clustering. Computer science review
34. Schafer, J. B., Konstan, J. A., and Riedl, J. (2001). E-commerce recommendation applications. Data mining and knowledge discover
35. Smith, B. and Linden, G. (2017). Two decades of recommender systems at amazon. com.
36. Goldberg, D., Nichols, D., Oki, B., and Terry, D. 1992. Using collaborative filtering to weave an information tapestry.
37. Goldberg, D., Nichols, D., Oki, B., and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. Communications of the ACM 35, 12, 61–70.
38. Bonhard, P. 2004. Improving recommender systems with social networking. In Pro- ceedings Addendum of the 2004 ACM Conference on Computer-Supported Cooperative Work. Chicago, IL, USA.
39. Shardanand, U. and Maes, P. 1995. Social information filtering: Algorithms for au- tomating “word of mouth”. In Proceedings of the ACM CHI Conference on Human Factors in Computing Systems. ACM Press, Denver, CO, USA, 210–217.

40.Herlocker, J., Konstan, J., and Riedl, J. 2002. An empirical analysis of de- sign choices in neighborhood-based collaborative filtering algorithms. Information Re- trieval 5, 4, 287–310.

41.Sowa, J. F. (1976). Conceptual graphs for a data base interface. 44.Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., and Wilkins, D. (2010). A comparison of a graph database and a relational database: a data provenance per- spective.

42.Mark Needham, A. E. H. (2019). Graph Algorithms: Practical Examples in Apache Spark Neo4, volume 1 of 1. O'Reilly, 1 edition. 46.Mojo, B. O. (2019). 2019 studio market share.