МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Рибак Олександр Борисович

УДК 004.4

Метод кластерного аналізу для прогнозування структур наночастинок

Галузь знань 12 «Інформаційні технології» за спеціальністю
122 «Комп'ютерні науки»
122 - БКР.А - 402.21930201

Автореферат
бакалаврської кваліфікаційної роботи на здобуття освітньої кваліфікації
«бакалавр з комп'ютерних наук»

Миколаїв – 2021

Бакалаврською кваліфікаційною роботою є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: д-р фіз-мат. наук, доцент, доцент кафедри комп'ютерної інженерії Лисенкво Едуард Анатолієвич

Рецензент: доцент кафедри інженерії програмного забезпечення Фісун Микола Тихонович

Захист відбудеться «25» червня 2021 р. о год. на засіданні екзаменаційної комісії (ауд. 2-406) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68 Десантників, 10.

З дипломною роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68 Десантників, 10.

Автореферат представлений «18» червня 2021 р.

Секретар
екзаменаційної комісії,
викладач кафедри ІІС А. С. Скакодуб

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Метою кластеризація наночастинок за допомогою комп'ютерного моделювання.

Об'єктом наночастинки та їх розташування на поверхні.

Предметом метод кластеризації Хошена-Копельмана.

Практичне значення розробленої системи полягає у тому, що вона дозволить моделювати та кластеризувати будь який масив данних, щодо розподілу наночастинок на поверхні

Дипломна робота складається зі вступу, 3 розділів, висновків, переліку джерел посилання та додатків. Загальний обсяг роботи складає 72 сторінок (без додатків), 10 рис., 1 додаток та 27 посилання на літературні джерела.

Ключові слова: наночастинки, нанотехнології, кластеризація, матриця, алгоритм Хошена-Копельмана.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі дипломної роботи обґрунтовано актуальність обраної теми, сформульовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У першому розділі проведено аналіз існуючих нанотехнологій та їх історію зародження у світі. Розглянуто загальну теорію наночастинок, проаналізовано переваги та недоліки різновидів наночастинок на перетині нашого життя.

У другому розділі бакалаврської роботи було проаналізовано існуючі методи кластеризації наночастинок, їх недоліки та переваги.

Було спроектовано систему кластеризації за алгоритмом Хошена компельмана.

Також було описано алгоритм Хошена-Копельмана, на основні матриці яка заповнена нулями та одиницями, де одиниці це наночастинки.

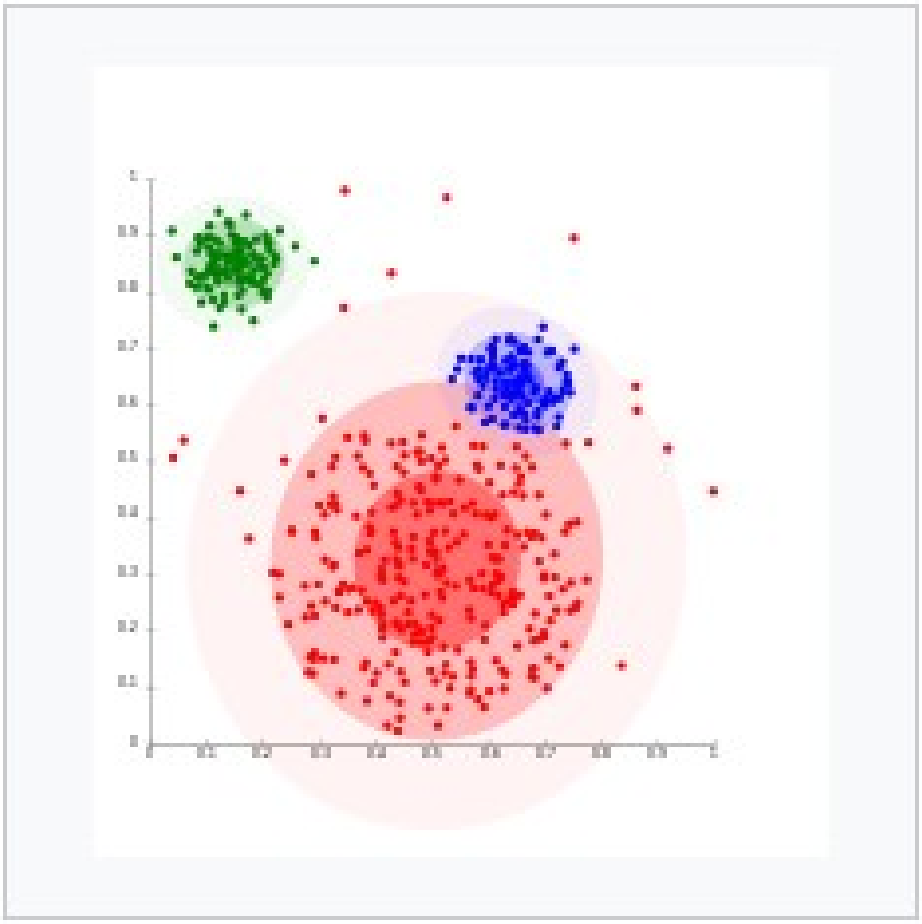

Рисунок 2.3 Приклад роботи алгоритма розподілу.

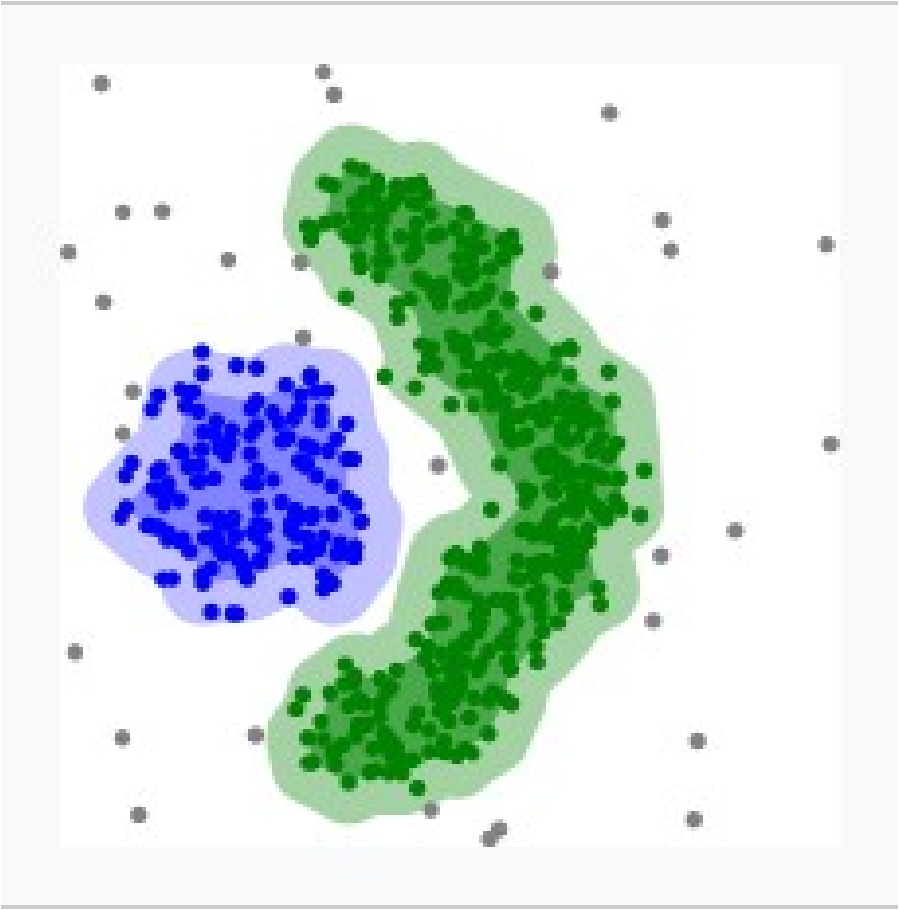

Рисунок 2.4 Приклад роботи Алгоритму щільності або DBSCAN.

Рисунок 2.5 Хід алгоритма середніх зсувів

Застосовано основні поняття теорії кластеризації наночастинок та комп’ютерного моделування – жорстка та м’яка кластеризації, моделювання ситуації за допомогою обчислювальних потужностей комп`ютера та кваліфікованої більшості.

Наприкінці розділу було описано різницю між класичним та модифікованим методами кластеризації.

В третьому розділі було виконано опис програмного забезпечення.

Програмний додаток написаний за допомогою мови комп'ютерного програмування C++. Розроблений додаток працює за допомого технології WindowsForm на відкритому сервері. Також, в розробленій інформаційній системі вирішені та показані проблеми алгоритму Хошена - Копельмана. Було вдосконалено його і тепер програма має можливість обробляти нескінченну кількість інформації і це залежить тільки від потужності копм'ютера.

У спеціальній частині роботи було виконано аналіз умов праці в комп'ютерній лабораторії вищого навчального закладу. Перевірено забезпечення вимог охорони праці. Виявлено, що оцінка умов праці на робочому місці відноситься до IV категорії, коли спостерігається робота у несприятливих умовах праці. З метою їх покращення розраховано припливно-витяжну вентиляцію повітропродуктивністю V ≥ 2500 м3/год. Підібрано вентилятор з необхідною витратно-напірною характеристикою.

В результаті виконання бакалаврської роботи було вдосконалено процес кластеризації наночастинок за рахунок створення системи повного жорсткого кластерного аналізу наночастинок на основі алгоритму Хошена-Копельмана.

Зазначену мету досягнуто завдяки виконання наступних завдань:

проаналізувано використання існуючих методів кластеризації наночастинок;

розроблено модифікований алгоритм кластеризації;

розроблено функціональну модель системи;

розроблено інформаційне та програмне забезпечення системи визначення кластерів на поверхні;

перевірено працездатність системи.

Описано функції та завдання

Описано функції та завдання, необхідні для розробки системи, основні поняття систем вибору на основі яких було створено веб програмний додаток, проведений розгляд методів Хошена-Копельмана для підрахунку загальної кількості голосів, кластерів наночастинок.

Застосовано основні поняття вибору кластерів та їх підрахунку в різних системах кластеризації жорстка та м'яка. У жорсткій кластеризації кожен об'єкт даних або повністю належить кластеру чи взагалі не належить. В м'якій кластеризації точка чи об'єкт даних може з певною ймовірністю належати більш ніж одному кластеру.

Проведено порівняльний аналіз вищезазначених моделей. В результаті даного порівняльного аналізу було виявлено, що найкращим методом підрахунку голосів буде модифікований метод Хошена - Копельмана створений для наночастинок розміром до 1 нанометра .

Практична значимість розробленої системи полягає у тому що вона є абсолютно жорсткою та результат залежить цілком від відсотка вказаного користувачем. У ній немає можливості фальсифікувати дані.

Термін фрактал (від латинського слова fractus – дробовий), був запропонований Бенуа Мандельбротом в 1975 році від Різдва Христова. Термін використовували для нерегулярних самоподібних математичних структур, вікористовуеться й надалі. Популярна сьогодні фрактальная геометрія отримала

свою назву лише в 1977 році завдяки його книзі «The Fractal Geometry of Nature». У роботах вченого використані наукові результати багатьох вчених, які працювали в цій же області (перш за все, Пуанкаре, Кантора, Хаусдорфа). Основне визначення фракталу Мандельброта було: "Фрактал – це структура, що складається з частин, які певним чином схожі на ціле".

Суворе визначення самоподібних множин було дано в 1981 р. Дж. Хатчинсоном. Він називав сукупність, схожу на себе, коли вона складається з декількох компонентів, подібних до всіх цих множин; Н. компонент отриманого афінського перетворення – обертання, стиснення, відображення вихідного речення. Однак самооцінка є необхідною, але недостатньою якістю фракталів. Адже неможливо дивитись на фрактальну точку або площину, намальовану клітинами.Головна особливість фракталів полягає в тому, що їх розмірність не вкладається в звичні геометричні уявлення. Для фракталів характерний геометричний «виріз». Звідси використовується спеціальна концепція фрактального виміру, введена Феліксом Хаусдорфом (1868-1942) та Абрама Самойловича Безіковича (1891-1970). У зв'язку з, щодо\ однакові числа вона дала ідеальним об’єктам класичної евклідової геометрії Значення, як відомо задовго до так званого топологічного виміру (іншими словами, нуль для точки, один – для плавної плавності Рядки, дві – для фігури та поверхні, три – для тіла та простору). Але узгоджується зі старим, топологічним, виміром ідеальних об'єктів, новим Вимір мав точнішу чутливість до кожного виду Недосконаліть реальних об’єктів, що дозволяють розрізнити і індивідуалізувати те, що раніше було безликим і нечітким. Так, відрізок прямої, відрізок синусоїди і найвигадливіший меандр невиразні з точки зору топологічної розмірності – всі вони мають топологічну розмірність, рівну одиниці, тоді як їх розмірність Хаусдорфа – Безиковича різна і дозволяє числом вимірювати ступінь звивистості. Розмірність фрактальних об'єктів не є є невід’ємною ознакою загальної геометрії. Однак у У більшості випадків фрактали нагадують предмети і щільно займають справжнє Космос, але не використовує його повною мірою. Нехай є безліч G в просторі Rn. Розіб'ємо простір Rn на n-мірні куби з довжиною ребра δ і позначимо число кубів, необхідних для покриття ними безлічі G, через N (δ). Тоді величина розмірності Хаусдорфа-Безиковича (званої також фрактальної розмірністю) D повинна задовольняти наступні умови:

Дане визначення можна спростити, зробивши його більш зручним для практичного застосування. Видно, що при $\delta \approx 0$, воно еквівалентно:

$$D \approx - \ln N(\delta) / \ln \delta.$$

Приклади абстрактних фракталів

Мандельброт запропонував не тільки визначення фракталів, але також і алгоритм побудови одного з них, що отримав назву на честь вченого. Алгоритм

побудови безлічі Мандельброта заснований на итеративном обчисленні за формулою:

$$Z [i + 1] = Z [i] * Z [i] + C,$$

де Z і C – комплексні змінні. Ітерації виконуються для кожної стартової точки C квадратної або прямокутної області – підмножині комплексній площині. Ітераційний процес продовжується до тих пір, поки Z [i] не вийде за межі кола заданого радіуса, центр якої лежить в точці (0,0), або після досить великої кількості ітерацій. Залежно від кількості ітерацій, протягом яких Z [i] залишається всередині кола, встановлюється колір точки C. Якщо Z [i] залишається в колі протягом досить великої кількості ітерацій, то ця точка растра забарвлюється в чорний колір. Безлічі Мандельброта належать саме ті точки, які протягом нескінченного числа ітерацій не йдуть в нескінченність.

Безліч Мандельброта

Так як кількість ітерацій відповідає номеру кольору, то точки, що знаходяться ближче до безлічі Мандельброта, мають більш яскравий колір.

Побудова іншого фрактального безлічі, сніжинки Коха, починається з правильного трикутника, довжина сторони якого дорівнює 1. Сторона трикутника вважається базовою ланкою для вихідного положення. Далі, на будь-якому кроці ітерації кожна ланка замінюється на який утворює елемент – ламану, що складається по краях з відрізків довжиною 1/3 від довжини ланки, між якими розміщуються дві сторони правильного трикутника зі стороною в 1/3 довжини ланки. Всі відрізки – сторони отриманої кривої вважаються базовими ланками для наступної ітерації. Крива, що отримується в результаті n-й ітерації при будь-якому доонечно n, називається предфракталом, і лише при n, що прагне до нескінченності, крива Коха стає фракталом. Отримується в результаті ітераційного процесу фрактальное безліч являє собою лінію нескінченної довжини, що обмежує кінцеву площу. Дійсно, при кожному кроці число сторін результуючого багатокутника збільшується в 4 рази, а довжина кожної сторони зменшується тільки в 3 рази, тобто довжина багатокутника на n-й ітерації дорівнює 3 * (4/3) n і прямує до нескінченності з ростом n.

Перші 5 поколінь сніжинки Коха

Площа під кривою, якщо прийняти площа утворює трикутника за 1, дорівнює:

У 80-х роках XX століття як простий засіб отримання фрактальних структур з'явився метод "Систем ітераційний Функцій" (Iterated Functions System – IFS). IFS являє собою систему функцій, що відображають одне багатовимірне безліч на інше. Найбільш проста реалізація IFS є аффінниє перетворення площині:

$$X '= A * X + B * Y + C$$

$$Y '= D * X + E * Y + F$$

У той же час американські вчені М. Барнслі і А. Слоан запропонували ідею стиснення і зберігання графічної інформації, засновану на міркуваннях теорії фракталів та динамічних систем. На підставі цієї ідеї був створений алгоритм фрактального стиснення інформації, що дозволяє стискати деякі зразки графічної інформації в 500-1000 разів. При цьому кожне зображення кодується кількома простими аффіннимі перетвореннями. Закодувавши якесь зображення двома аффіннимі перетвореннями, воно однозначно визначається за допомогою 12-ти коефіцієнтів. Якщо визначити початкову точку ітераційного процесу (наприклад, X = 0 Y = 0) і запустити цей процес, то через кілька ітерацій сукупність отриманих точок буде описувати закодоване зображення.

Як приклад використання IFS для побудови фрактальних структур, можна привести криву "дракона" Хартера-Хейтуея [5].

"Дракон" Хартера-Хейтуея

Використання IFS для стиснення звичайних зображень, таких як фотографії засноване на виявленні локального самоподібності (на відміну від фракталів, де спостерігається глобальне самоподоба). За алгоритмом Барнслі відбувається виділення в зображенні пар областей, найменша з яких подібна до більшої, і збереження декількох коефіцієнтів, що кодують перетворення, що переводить більшу область в меншу. Потрібно, щоб безліч таких областей покривало все зображення. Відновлювальний алгоритм повинен застосовувати кожне перетворення до деякій підмножині, що належить області, відповідної застосовується перетворення.

Фрактали з великою точністю описують багато фізичних явищ і природні утворення: хмари, турбулентні течії, гілки дерев, кровоносні судини. Мандельброт свого часу зауважив: "Чому геометрію часто згадують холодний і сухий? Однією з причин є його нездатність описати форму Хмари, гори, дерева або узбережжя. Хмари – не сфери, гори – НЕ конуси, узбережжя – не кола, а кора не гладка і блискавка не поширюється Прямий. Природа показує нам не тільки вищий ступінь, але і зовсім інший рівень складності.

У машинній графіці фрактальні підходи приходять на допомогу, наприклад, коли потрібно, за допомогою декількох коефіцієнтів, задати лінії і поверхні дуже форми. Фрактальна геометрія сьогодні незамінна при комп'ютерній генерації хмар, гір, поверхні моря, інших складних «неевклідових» об'єктів, образи яких нагадують природні.

3) Фрактали в природі

У реальному житті фрактальні об'єкти мають цілком певні межі фрактальности, в тому числі і самоподібності. Проте, фрактали – це дуже зручна і наочна абстракція, яка сьогодні вже широко застосовується при моделюванні природних процесів. При цьому спектр застосування фракталів постійно

розширюється, сьогодні він застосовується і до моделювання інформаційного простору.

Один з кращих прикладів прояви фракталів в природі -Структура берегових ліній. Дійсно, дивиться на кілометровий відрізок узбережжя зрізаний як на сто кілометрів. Досвід показує, що довжина берегової лінії L залежить від масштабу l, що проводиться вимірювання, і збільшується зі зменшенням останнього за степеневим законом $L = \Lambda\ l1\text{-}\alpha$, Λ = const. Так, наприклад, для узбережжя Великобританії $\alpha \approx 1.24$, тобто, так звана, фрактальна розмірність берегової лінії Великобританії дорівнює 1.24.

Берегова лінія узбережжя Великобританії

Нещодавно Б. Саповаль з Політехнічної школи в Палезо (Франція) і його колеги створили комп'ютерну модель ерозії узбережжя. У моделі речовина руйнувалося або під прямим впливом хвиль, або повільним "вивітрюванням", коли мінерали розчинялися у воді. Узбережжя було розділене на рівні ділянки, причому в моделі типи каменів на цих ділянках вибиралися випадковим чином. Ерозійна сила моря залежить від того, наскільки сильно заглушуються хвилі. У вузькому затоці або бухті вода завжди спокійніше. Саповаль припустив, що глушіння хвиль посилюється в міру того, як берег стає більш порізаним. Модель показала, що спочатку гладка берегова лінія стрімко набуває нерівний профіль з виступами і безліччю відокремлених від берега островів. При моделюванні берегових ліній використовувалися двовимірні стохастичні фрактали, які виходять в тому випадку, якщо в ітераційне процесі випадковим чином варіювати деякі його параметри. Утворився при моделюванні берег дуже нагадував Східне узбережжя США.

У 2004 році в Ньюфаундленді біологом Гі Нарбонна з університету Кінгстона (Канада) була відкрита рідкісна викопна природна структура фрактального типу. Були знайдені сліди організмів, що жили на Землі близько 575 мільйонів років тому, і не належали ні до рослин, ні до тварин, і називаються рангеоморфамі. Вони були нездатні рухатися і не мали репродуктивних органів, а розмножувалися, створюючи нові відгалуження. Організми збиралися у фрактальні структури з розгалужуються частин. Як з'ясувалося, кожен розгалужених елемент фрактальних структур складався їх безлічі трубок, утримуваних разом напівтвердим органічним скелетом організмів. Нарбонн виявив рангеоморфи, зібрані в кілька різних форм. Фрактальний малюнок представляється досить складним, але, за словами дослідника, схожість організмів один з одним забезпечувалося досить простим геном.

Приблизно півстоліття в біології відомий закон, який говорить про це

багато властивостей організмів, від тривалості життя до кількості молодих до

Швидкість обміну речовин пропорційна масі тіла в градусах n / 4, де n – ціле число. У той же час природа закону залишалася загадкою понад півстоліття. На

перший погляд, це має бути три замість чотирьох, бо натовп пропорційна кубу розміром тіла. Кілька років тому пояснення, було знайдено. Справа в тому, що пронизують кожен організм мережі – кровоносна у тварин або капілярна у рослин – володіють властивостями фракталів. Фрактальность цих мереж як раз і призводить до додавання ще одного "вимірювання" у живих організмів.

І нарешті, весь Всесвіт, відповідно до гіпотези російського фізика С. Хайтун, є фракталом, причому єдиним відомим в природі, повністю задовольняє класичним визначенням. У фізиці відомий факт, що щільність космічних об'єктів стрімко падає з їх розмірами. Ще в 50-х роках радянські фізики-теоретики прийшли до висновку, що "нескінченна" щільність Всесвіту дорівнює нулю. Ця ідея і нові уявлення про фрактальности Всесвіту підтверджують один одного. Справа в тому, що щільність всякого фрактала, розташованого в тривимірному просторі, тотожно дорівнює нулю. Класичні фрактали мають "усюди порожній" структурою, яка при проникненні в неї "розступається" до нескінченності. Разом з тим, реальні системи нескінченного поглиблення в свою структуру не дозволяють; на якомусь кінцевому етапі структура, будь то, скажімо, сніжинка або кровоносна система людини, втрачає свій "фрактальний" вид – реальні структури лише "фракталоподобни". Відповідно до гіпотези Хайтун, дозволяючи – через свою нескінченності – нескінченне проникнення в свою структуру, Всесвіт, на думку багатьох дослідників, є єдиним "справжнім" фракталом, маючи нульову нескінченну щільність.

В даний час інформаційний простір в цілому, з огляду на його обсягів і динаміки зміни, прийнято розглядати як е. З багатьма моделямиІнформаційний простір досліджує структурні зв’язки між тематичними Набори входять у цю кімнату. Водночас кількісні показники цьогоНабори підкоряються гіперболічному закону (з можливими статичнимиЗміни). Сьогодні все частіше в моделюванні інформаційного просторувикористовується фрактальний підхід, заснований на властивостяхСамоподібність інформаційного простору, d. Н. збереження інтер’єру Структури множин із зміною їх розмірів або масштабами спостереження ззовні. Самоподібність інформаційного простору в основному виражається втому що з його лавиноподібним зростанням за останні кілька десятиліть частота іРейтинг у таких областях, як джерела, автори, теми авторівпрактично не змінюють форму. Звідси застосування фрактальної теорії в РосіїАналіз інформаційного простору дозволяє поглянути на спільну позиціюЗакони, що складають основу інформатики. Наприклад тематичноІнформаційні поля представляють сьогодні країни, що розвиваються самоподібні структури, які по суті є стохастичними фракталами, такяк їх самооцінка застосовується лише на рівні математичних очікувань, наприклад розподіл кластерів за розміром. В інформаційному просторі виникають, формуються, зростають іКластери – групи пов’язаних документів – множаться. Системи,, засновані на кластерном аналізі, самостійно виявляють нові ознаки

об'єктів і розподіляють об'єкти за новими групами. Не так давно в Інтернет з'явився сервіс Touchgraph (www.touchgraph.com), який наочно демонструє появу кластерних утворень, сформованих подобою інформаційних об'єктів, зокрема, Web-сайтів (Touchgraph Google Browser). Нижче представлений приклад такої візуалізації:Об'єднання Web-сайтів за ознакою подібності Чим же визначається природа фрактальної структури інформаційного простору, що породжується такими кластерними структурами? З одного боку, параметрами рангових розподілів, а, з іншого боку, документальних масивів, визначенні взаємопов'язаних груп документів, для спрощення процесу перегляду при пошуку необхідної інформації, знаходження унікальних документів з колекції, виявлення дублікатів або близьких за змістом документів.

Фрактальний принцип самоподібності передбачає нескінченне дроблення набору об'єктів зі збереженням їх властивостей. У тематичних інформаційних потоках, наприклад, можна спостерігати подібність сюжетних ланцюжків, одержуваних при уточненні запиту (звичайно в певних рамках). Разом з тим, сьогодні багатьма дослідниками розглядається не дроблення, а природне зростання розмірів інформаційного простору.

Властивості самоподібності фрагментів інформаційного простору наочно демонструє новий інтерфейс представлений на веб-сайті служби News Is Free (http://newsisfree.com). На цьому сайті можна переглянути статус інформаційного простору у вигляді посилань на джерела та окремі повідомлення. При цьому враховується два основних параметри відображення – ранг популярності і «свіжість» інформації. В рамках цієї моделі можна спостерігати «дроблення» груп джерел при збільшенні рангу популярності і «свіжості» видань. Коли цей ранг стає досить високим, дроблення не дозволяє без особливих зусиль читати назви джерел і ідентифікувати окремі документи.

Web-простір, будучи, мабуть, найдинамічнішою частиною інформаційного простору, характеризується великій кількості прихованих в ньому неявних експертних оцінок, реалізованих у вигляді гіперпосилань. У листопаді 1999 року один з керівників інституту пошуку та аналізу текстів, що входить в дослідницький підрозділ IBM, Андрій Брёдер (Andrei Broder) і його співавтори з компаній AltaVista, IBM і Compaq математично описали "карту" ресурсів і гіперсвязів.

Кластери публікацій служби News Is Free

Дослідження спростували поширену думку, ніби Internet – це єдине густе простір. Відстеження за допомогою пошукової системи AltaVista закінчено

200 мільйонів веб-сторінок і мільярди посилань, розміщених на них

орієнтований графік, у якому вершини відповідають веб-сторінкам і ребрам – з'єднує ці сторінки гіперпосиланнями. В рамках цієї моделі завдання аналізу полягає була знайдена структура посилань між окремими веб-сторінками:

центральне ядро (28% веб-сайтів) – сильні компоненти зв'язку (SCC). 22% веб-сайтів – це "Початкові веб-сторінки" (IN). вони включають Гіперпосилання, які в підсумку ведуть до ядра, але від ядра до них не можна туди їхати.

стільки ж – 22% – "Кінцеві веб-сторінки" (OUT), до яких можна отримати доступ Посилання з ядра, але повернутися назад не можна. 22% веб-сторінок – вкладень – повністю ізольованих від центрального ядра:

ці або "миси" посилаються на інші сайти гіперпосиланнями

Категорії або "перешийок", що з'єднують дві веб-сторінки, яких немає

Модель також враховує "острови", які не перекриваються з іншими

Інтернет-ресурси. Єдиний спосіб визначити ресурси цієї групи – це знати адресу. Топологія та властивості моделі були приблизно однаковими для

різні підмножини веб-простору, що підтверджує спостереження

що "павутина – це фрактал", тобто властивості будови всього мухи веб-простору, також відповідає дійсності

Алгоритми, що використовують інформацію про структуру Web-простору, імовірно має працювати і на окремих його подмножествах. Інформація про структуру Web-простору вже досить широко використовується при вирішенні багатьох завдань, наприклад, для оптимізації ефективності механізмів сканування, при побудові нових Web-сервісів, для вирішення завдань аналізу і прогнозу.

5) Фрактали і тимчасові ряди

Новинна складова інформаційного простору Інтернет сьогодні настільки значна за своїми обсягом і динаміці, що може розглядатися як потужний інформаційний потік. Причому потік досить неоднорідний, який може характеризуватися великою кількістю параметрів, серед яких виділяються такі, як джерела інформації (веб-сайт) та тематики. Саме їх можна розглядати, як лежать на поверхні основи для кластеризації.

У той час, як для традиційних засобів наукової комунікації підходи до кластеризації з точки зору теорії фракталів були вперше досліджені Ван Рааном, які аналізували масиви статей і зв'язку, утворені цитуванням, інформаційні потоки повідомлень з Інтернет до останнього часу не асоціювалися з фракталами, що пов'язано з проблемами ідентифікації інформаційних потоків як фрактальних множин, а також з труднощами знаходження основ для побудови кластерів – повідомлень в політематичних потоках, що породжують багаторазове цитування.

З цієї ж причини досліджуються кількісні характеристики лише тематичних інформаційних потоків, які характеризуються ітеративна при формуванні і цілком доступні як для кількісного, так і для якісного аналізу.

Обсяги повідомлень в тематичних інформаційних потоках утворюють тимчасові ряди. Тимчасові ряди, породжувані тематичними інформаційними

потоками, також мають фрактальні властивості і можуть розглядатися як стохастичні фрактали. Цей підхід розширює сферу застосування теорії фракталів на інформаційні потоки, динаміка яких описується засобами теорії випадкових процесів.

З іншого боку, теорія фракталів розглядається як підхід до статистичного дослідження, який дозволяє отримувати важливі характеристики інформаційних потоків, не вдаючись у детальний аналіз їх внутрішньої структури та зв'язків. Одним з основних властивостей фракталів є самоподібність (скейлинг). Як показано в роботах С.А. Іванова, для послідовності повідомлень тематичних інформаційних потоків відповідно до скейлінговим принципом, кількість повідомлень, резонансів на події реального світу пропорційно деякій мірі кількості джерел інформації (кластерів) і ітераційно триває протягом певного часу. Точно так же, як і в традиційних наукових комунікаціях, зростаюче безліч повідомлень в Інтернет по одній тематиці в часі являє собою динамічну кластерну систему, яка виникає в результаті ітераційних процесів. Цей процес пояснюється ре публікування, прямий або спільної цитованістю, різними публікаціями – відображеннями одних і тих же подій реального світу, прямими посиланнями і т.д. Крім того, для більшості тематичних інформаційних потоків спостерігається збільшення їх обсягів, причому на коротких тимчасових інтервалах – лінійний зростання, а на тривалих – експонентний.

Фрактальна розмірність в кластерної системі, відповідної тематичним інформаціонним потокам, показує ступінь заповнення інформаційного простору повідомлень протягом певного часу.

Показник Херста

Сьогодні у зв'язку з розвитком теорії стохастичних фракталів стає популярною така характеристика часових рядів як, показник Херста (H). Відомо, що він пов'язаний з традиційною «клітинної» фрактальної розмірністю (D) простим співвідношенням:

$D + H = 2$.

Умова, при якому показник Херста пов'язаний з фрактальної «клітинної» розмірністю, визначено Е. Федер наступним чином: «... розглядають клітини, розміри яких малі в порівнянні як з тривалістю процесу, так і з діапазоном зміни функції; тому співвідношення справедливо, коли структура кривої, що описує фрактальну функцію, досліджується з високою роздільною здатністю, механізмом розвитку інформаційних кластерів, який відображає природу інформаційного простору. Поява нових публікацій збільшує розмірність вже існуючих кластерів і є причиною утворення нових.

Фрактальні властивості характерні для кластерів інформаційних Web-сайтів, на яких публікуються документи, що відповідають певним тематикам. Ці кластери,

як набори тематичних документів, представляють собою фрактальні структури, які мають низку унікальних властивостей. Наприклад, російськими дослідниками (С. Іванов та ін.), Визначена фрактальная розмірність подібних інформаційних масивів, що змінюється в межах від 1.05 до 1.50, що свідчить про невелику щільності заповнення кластерів документами по одній темі. Як один з основних законів відображають самоподоба інформаційного простору можна назвати закон Зіпфа. У 1949 році професор філології з Гарварда Дж. Зіпф зібрав достатній статистичний матеріал, і експериментально показав, що розподіл слів природної мови підпорядковується закону: "Якщо до якогось досить великого тексту скласти список всіх зустрілися в ньому слів, а потім ранжувати ці слова, тобто розташувати їх у порядку убування частоти народження в даному тексті і пронумерувати в порядку зростання, то для будь-якого слова твір його порядкового номера (рангу) цьому списку і частоти його народження в тексті буде величиною постійною. " Вчений описав виявлену ним закономірність розподілу слів в текстах англійською мовою:

невелика кількість слів, таких як "the", "and" в англійській мові, які мають дуже високий ранг;

середня кількість слів має середній ранг;

велика кількість слів має дуже низький ранг.

Таким чином: f * r = c, де f – частота народження слова в тексті; r – ранг (порядковий номер) слова в списку; з – емпірична постійна величина. Так, наприклад, для англійських текстів константа Зіпфа дорівнює приблизно 0,1. Для російської та української мов коефіцієнти Зіпфа складаю приблизно 0,06-0,07.

Існують також закономірності, відкриті іншими вченими (перш за все, Бредфордом – для періодичних видань і Лотки – для розподілу авторів), які є уточнюючими наслідками закономірностей Зіпфа, і також свідчать про самоподобу інформаційного простору.

Теорія фракталів тісно пов'язана з кластерним аналізом, вирішальним завдання виділення компактних груп об'єктів з близькими властивостями. Кластеризація сьогодні застосовується при реферування великих тобто в локальному межі ». Ще однією важливою умовою є самоаффінность функції. Не вдаючись в подробиці зауважимо, що для інформаційних потоків це властивість інтерпретується як самоподоба, що виникає в результаті процесів їх формування. Можна помітити, що зазначеними властивостями володіють не всі інформаційні потоки, а лише ті, які характеризуються достатньою потужністю і ітеративний при формуванні. При цьому тимчасові ряди, побудовані на підставі потужних тематичних інформаційних потоків, цілком задовольняють цій умові. Тому при розрахунку показника Херста, фактично визначається і такий показник тематичного інформаційного потоку, як фрактальна розмірність.

Відомо, що показник Херста представляє собою міру персистентності - схильність процесу до трендам (на відміну від звичайного броунівського руху). Значення H> ½ означає, що спрямована в певний бік динаміка процесу в минулому, найімовірніше, спричинить продовження руху в тому ж напрямку. Якщо H <½, то прогнозується, що процес змінить спрямованість. H = ½ означає невизначеність – броунівський рух.

Для вивчення фрактальних характеристик тематичних інформаційних потоків вивчалися значення показника Херста за певний період для часових рядів, складених з кількості які належать до них. Показник Херста пов'язують з коефіцієнтом нормованого розмаху (R / S), де R – обчислюється певним чином «розмах» відповідного часового ряду, а S – стандартне відхилення.

Показник Херста обчислюється за наступним алгоритмом. Спочатку обчислюється середнє значення вимірюваної змінної (в нашому випадку кількість повідомлень в інформаційному потоці) за N днів:

Потім розраховується накопичилася відхилення ряду вимірювань (t) від середнього:

Після цього розраховується різниця максимального і мінімального накопичився відхилення, яка і називається "розмахом":

Стандартне відхилення розраховується за відомою формулою:

Свого часу Херст експериментально виявив, що для багатьох часових рядів справедливо:

Саме коефіцієнт H і отримав назву показника Херста.

Опис обчислювального експерименту

В якості експериментальної бази для дослідження фрактальних властивостей тематичних інформаційних потоків використовувалася система контент-моніторингу InfoStream, розроблена в Інформаційному центрі «Електронні вісті». Ця система, яка застосовується для вирішення завдань автоматизованого збору новинної інформації з відкритих Web-сайтів і забезпечення доступу до неї в пошукових режимах, в даний час охоплює понад 2000 джерел інформації – більш 40000 унікальних новинних повідомлень на добу. У ретроспективних базах даних системи накопичено понад 25 млн. Повідомлень.

Тематика досліджуваного інформаційного потоку визначалася запитом до системи InfoStream, що складається всього з одного слова «Microsoft». Ретроспективний період дослідження становив весь 2005 рік і 2 місяці 2006 року, тобто 424 дня (N = 424). В результаті пошуку було знайдено 42357 релевантних документів.

Вихідні дані були отримані з інтерфейсу режиму «Динаміка появи понять». На підставі обробки цих даних була отримана повна картина експериментальних даних – часовий ряд за вказаний період.

Фрагмент діаграми динаміки народження поняття «Microsoft»

Для цього часового ряду за формулою (6) булообчислено стандартне відхилення (S = 43.71). Одночасно, за допомогою механізму формування основних сюжетів, що входить до складу системи InfoStream, були визначені основні події, що призвели до виникнення пікових значень на діаграмі.

Динаміка накопичення відхилення дозволила визначити «розмах» цього параметра (R = 1207.64).

І нарешті, для значення N = 424 був обчислений показник Херста, який виявився рівним 0,62, що свідчить про позитивну персистентності всього тимчасового ряду.

Крім того, були виконані розрахунки показників Херста для всіх значень N, починаючи з 5.

Вивчення такої характеристики, як показник Херста дозволяє прогнозувати динаміку інформаційних потоків, повідомлення яких відображають процеси, що відбуваються в реальному світі.

Наведені в прикладі дані підтвердили лежить в основі дослідження припущення про ітеративності процесів в інформаційному просторі. Передруку, цитування, прямі посилання і т.п. породжують самоподоба, що виявляється в стійких статистичних розподілах і відомих емпіричних законах. Скейлінговий принцип пояснюється також схожістю ментальності авторів, що публікують повідомлення в Інтернет. Разом з тим різні маркетингові, рекламні, PR-кампанії ведуть до стрибкоподібним змінам в стабільних статистичних закономірностях, різких перепадів і спотворень в порівнянні зі стандартними статистичними розподілами.

В результаті експерименту також підтверджено наявність статистичної кореляції в інформаційних потоках на тривалих тимчасових інтервалах.

Зокрема, на даному прикладі, показана персистентність процесу, що говорить, про загальну середню збільшенні публікації про компанії Microsoft, періодичному появі піків, пов'язаних, як правило, з двома підтемами-кластерами – особистістю Білла Гейтса (чотири з п'яти топ-кластерів) і відображеннями вірусних атак (п'ятий топ-кластер).

АНОТАЦІЯ

Бакалаврської кваліфікаційної роботи
студента 402 групи
ЧНУ Ім. Петра Могили
Рибака Олександра Борисовича

Назва роботи: Фрактальний аналіз різних зображень кластерів наночастинок методом box-counting

Мета та основні результати роботи: створення програмного забезпечення для аналізу кластерів наночастинок для їх ідентефікації та знаходження. Користувач матиме можливість ідентифікувати деякий кластер наноелментів, знайти кластер у конкретному матеріалі та відрізнити різні кластери за їх властивостями.

Ключові слова: сайт – каталог, комп`ютерні та відеоігри, критерії, інформація та дані, онлайн каталог.

ABSTRACT

Bachelor's qualification work

student of 402 group

Petro Mohyla Black Sea National University

Ribaka Oleksandra Borisovicha

Title: Creating information - reference catalog of games with a web - interface.

Purpose and main results of the work: creation of an information - reference catalog of games with a web - interface and realization of a search system for information about a particular game, due to certain criteria and data that are in the catalog itself; site - catalog of games, which will systematize information and data about computer and video games due to certain criteria that will help with the search in the catalog.

Keywords: site – catalog, computer and video games, criteria, information and data, online catalog.