УДК 004.652

DOI 10.31651/2076-5886-2022-1-62-72

PACS 07.05.Kf

КУЦЕНКО Олександр Анатолійович
аспірант спеціальності «Прикладна математика», Черкаський національний університет ім. Б. Хмельницького
e-mail: alexkutsenko1995@gmail.com
ORCID 0000-0002-6220-6034

ПІСКУН Олександр Варфоломійович
кандидат технічних наук, доцент, завідувач кафедри прикладної математики та інформатики, Черкаський національний університет ім. Б. Хмельницького
e-mail: piskun@ukr.net
ORCID 0000-0001-5334-6337

МОДЕЛІ ОРГАНІЗАЦІЇ ДАНИХ ТА ЇХ ОПТИМІЗАЦІЯ ПРИ ПРОВЕДЕННІ СЕМАНТИЧНОГО АНАЛІЗУ КОНТЕНТУ СОЦІАЛЬНИХ МЕРЕЖ

У роботі досліджено фундаментальні підходи до організації даних в інформаційних системах та їх застосування для оптимізації семантичного аналізу контенту соціальних мереж. Проаналізовано класичні моделі даних – ієрархічну, мережеву та реляційну, визначено їх переваги та обмеження у контексті обробки великих обсягів текстової інформації. Розглянуто методологію контент-аналізу як інструменту дослідження соціальних комунікацій та семантичний аналіз тексту як етап автоматичного розуміння природної мови. Встановлено взаємозв'язок між вибором моделі організації даних та ефективністю алгоритмів семантичної обробки контенту соціальних мереж.

Ключові слова: *моделі даних, організація даних, контент-аналіз, семантичний аналіз тексту, соціальні мережі, оптимізація.*

Вступ

Сучасний етап розвитку інформаційних технологій характеризується експоненційним зростанням обсягів цифрової інформації, що циркулює в мережі Інтернет [1]. Особливо інтенсивно цей процес відбувається у соціальних мережах, які стали основним каналом комунікації для мільярдів користувачів. За оцінками дослідників, щодня у соціальних мережах генерується понад 500 мільйонів повідомлень, що створює як нові можливості, так і виклики для аналізу цієї інформації [2].

Ефективна обробка таких масивів даних потребує не лише потужних обчислювальних ресурсів, але й оптимальної організації самих даних. Вибір моделі організації даних безпосередньо впливає на швидкість доступу до інформації, складність алгоритмів обробки та якість кінцевих результатів аналізу [3]. Водночас зростає потреба у автоматизованих системах аналізу контенту, здатних виявляти приховані закономірності, тенденції та смислові зв'язки у великих масивах текстових даних.

Контент-аналіз як метод дослідження зародився у соціології та журналістиці, але з розвитком обчислювальної техніки трансформувався у потужний інструмент автоматичної обробки текстів [4]. Семантичний аналіз, що є складовою контент-аналізу, дозволяє виявляти не лише явний, але й прихований зміст повідомлень, що

особливо важливо для розуміння суспільних настроїв, виявлення тенденцій та прогнозування поведінки користувачів соціальних мереж [5].

Метою даного дослідження є аналіз класичних моделей організації даних та визначення їх оптимальності для задач семантичного аналізу контенту соціальних мереж.

Актуальність роботи обумовлена необхідністю розробки ефективних підходів до обробки великих обсягів неструктурованої текстової інформації у режимі реального часу.

Виклад основного матеріалу

1. Моделі організації даних: теоретичні основи

1.1. Поняття даних в інформаційних системах

У контексті інформаційних систем дані розуміються як формалізоване представлення інформації, придатне для зберігання, передачі та автоматизованої обробки [6]. На відміну від інформації, яка має семантичне значення для користувача, дані є нейтральною формою представлення фактів, що може піддаватися багаторазовій інтерпретації залежно від контексту використання [7].

Модель даних визначає логічну структуру організації інформації у базі даних, специфікує операції над даними та обмеження цілісності. Історично склалися три класичні моделі організації даних, кожна з яких відображає певний рівень абстракції та підходів до структурування інформації [8]. Ієрархічна модель виникла першою у 1960-х роках, за нею послідували мережева та реляційна моделі. Вибір конкретної моделі залежить від специфіки предметної області, характеру зв'язків між даними та вимог до швидкодії системи.

1.2. Ієрархічна модель даних

Ієрархічна модель представляє дані у вигляді деревоподібної структури, де кожен елемент має не більше одного батьківського вузла [9]. Така організація природно відображає відносини підпорядкування та успадкування, що робить модель інтуїтивно зрозумілою для користувачів. Основними поняттями ієрархічної моделі є рівень, вузол та зв'язок. Кореневий вузол розташований на найвищому рівні і не має батьківських елементів, тоді як усі інші вузли мають рівно один батьківський вузол і можуть мати довільну кількість дочірніх елементів.

Кожен вузол ієрархічної структури представляє собою сегмент даних, що містить набір атрибутів конкретного об'єкта предметної області. Доступ до даних здійснюється шляхом проходження ієрархічного шляху від кореневого вузла до цільового елемента, що забезпечує передбачувану швидкість операцій читання [10]. Ієрархічна модель характеризується ефективним використанням пам'яті та високою швидкодією завдяки фізичній близькості пов'язаних даних на носії інформації.

Проте ієрархічна модель має суттєві обмеження. По-перше, вона не підходить для представлення складних зв'язків типу "багато-до-багатьох", що часто зустрічаються у реальних задачах [11]. По-друге, будь-яка зміна структури даних вимагає повної перебудови бази даних, що робить систему негнучкою. По-третє, для доступу до даних необхідно знати повний ієрархічний шлях, що ускладнює формування довільних запитів. Ці недоліки обмежують застосування ієрархічної моделі для аналізу контенту соціальних мереж, де зв'язки між об'єктами мають складний та динамічний характер.

1.3. Мережева модель даних

Мережева модель виникла як розширення ієрархічної і дозволяє представляти складніші структури даних шляхом допущення множинних батьківських зв'язків [12]. У мережевій моделі дані організовані у вигляді графа, де вузли представляють об'єкти, а ребра відображають зв'язки між ними. Така організація дозволяє природно моделювати відносини типу "багато-до-багатьох" без дублювання інформації.

Основними конструкціями мережевої моделі є запис та набір. Запис являє собою іменовану сукупність атрибутів, що описують деякий об'єкт предметної області. Набір визначає зв'язок типу "один-до-багатьох" між двома типами записів, де один запис виступає власником, а інші є членами набору [13]. Фізична реалізація зв'язків здійснюється за допомогою покажчиків, що забезпечує ефективну навігацію по структурі даних. Важливою особливістю мережевої моделі є можливість переміщення як від власника до членів набору, так і у зворотному напрямку.

Мережева модель забезпечує більшу гнучкість порівняно з ієрархічною, дозволяючи представляти довільні зв'язки між об'єктами. Швидкодія мережевих баз даних залишається високою завдяки прямому доступу через покажчики [14]. Однак складність структури призводить до того, що розробка та супровід мережевих баз даних вимагає високої кваліфікації. Крім того, як і в ієрархічній моделі, для доступу до даних необхідно знати структуру зв'язків, що робить систему менш зручною для кінцевих користувачів. Зміна структури бази даних залишається складною задачею, що обмежує адаптивність системи до змінних вимог.

1.4. Реляційна модель даних

Реляційна модель, запропонована Едгаром Коддом у 1970 році, здійснила революцію в організації баз даних завдяки простоті концепції та потужному математичному апарату [15]. У реляційній моделі дані представлені у вигляді таблиць, де рядки відповідають записам, а стовпці атрибутам об'єктів. Така форма представлення інтуїтивно зрозуміла користувачам і не вимагає знання складних ієрархічних чи мережевих структур.

Математичною основою реляційної моделі є теорія множин та реляційна алгебра, що дозволяє формально визначити операції над даними [16]. Кожна таблиця представляє відношення, яке є підмножиною декартового добутку доменів її атрибутів. Основними поняттями моделі є атрибут як іменований стовпець відношення, кортеж як рядок таблиці, первинний ключ як унікальний ідентифікатор запису, та зовнішній ключ для встановлення зв'язків між таблицями. Реляційна модель вводить строгі обмеження цілісності, зокрема цілісність сутностей та посилальну цілісність, що гарантує коректність даних.

Перевагами реляційної моделі є незалежність даних від прикладних програм, що полегшує модифікацію структури бази без зміни програмного коду [17]. Декларативна мова запитів SQL дозволяє формулювати довільні запити без необхідності специфікації процедури доступу до даних. Реляційна модель забезпечує високий рівень захисту даних та підтримку паралельного доступу багатьох користувачів. Проте реляційні системи поступаються ієрархічним та мережевим за швидкодією операцій читання, що критично для деяких застосувань. Крім того, відображення складних об'єктів з множинними зв'язками може вимагати багатьох таблиць та операцій з'єднання, що знижує ефективність.

1.5. Сучасні моделі даних

Розвиток інформаційних технологій призвів до появи нових моделей даних, орієнтованих на специфічні застосування. Об'єктно-орієнтована модель інтегрує концепції об'єктно-орієнтованого програмування з організацією даних, дозволяючи зберігати складні об'єкти з методами та успадкуванням [18]. Багатовимірна модель оптимізована для аналітичної обробки великих масивів агрегованих даних у системах підтримки прийняття рішень. Документо-орієнтовані бази даних, що набули популярності з появою NoSQL руху, зберігають дані у вигляді самоописуваних документів без жорсткої схеми, що забезпечує гнучкість та горизонтальну масштабованість [19].

Для обробки даних соціальних мереж особливий інтерес представляють графові бази даних, які природно моделюють зв'язки між користувачами та контентом [20]. У графовій моделі вузли представляють сутності, а ребра зв'язки між ними, причому як вузли, так і ребра можуть мати довільні атрибути. Такий підхід оптимізує операції обходу графа та пошуку шляхів, що критично для аналізу соціальних структур. Гібридні підходи поєднують переваги різних моделей, використовуючи реляційні бази для структурованих даних та NoSQL рішення для неструктурованого контенту.

2. Контент-аналіз як метод дослідження

2.1. Методологічні основи контент-аналізу

Контент-аналіз визначається як систематичний, об'єктивний та кількісний метод аналізу змісту комунікацій [21]. Виникнувши у соціології для вивчення засобів масової інформації, цей метод згодом знайшов застосування у психології, політології, маркетингу та інших галузях. Сутність контент-аналізу полягає у перетворенні якісної текстової інформації у кількісні показники шляхом виявлення та підрахунку частоти появи певних елементів змісту [22].

Об'єктивність контент-аналізу досягається завдяки чітко визначеним процедурам та критеріям класифікації матеріалу. Дослідник формулює категорії аналізу, що відображають предмет дослідження, та одиниці аналізу, тобто елементи тексту, які підлягають ідентифікації та підрахунку [23]. Систематичність означає, що аналіз охоплює весь досліджуваний матеріал згідно з певними правилами вибірки. Кількісний характер аналізу дозволяє застосовувати статистичні методи обробки результатів та перевіряти гіпотези про зв'язки між характеристиками контенту та іншими змінними.

2.2. Етапи проведення контент-аналізу

Реалізація контент-аналізу передбачає послідовне виконання кількох етапів, кожен з яких має методологічне значення. На першому етапі визначається сукупність джерел інформації згідно з критеріями релевантності дослідженню [24]. Для аналізу соціальних мереж це можуть бути публічні профілі користувачів, групи за інтересами, коментарі до публікацій тощо. Важливо забезпечити репрезентативність вибірки, що може вимагати стратифікації за часовими періодами, географічним розташуванням або демографічними характеристиками авторів.

Другий етап передбачає формування вибіркової сукупності повідомлень, якщо повний аналіз генеральної сукупності неможливий або недоцільний. На третьому етапі виявляються одиниці аналізу, якими можуть бути окремі слова, словосполучення, тематичні блоки або цілі документи залежно від мети дослідження [25]. Четвертий етап

полягає у визначенні одиниць підрахунку, наприклад частоти появи категорії, обсягу тексту, присвяченого темі, або інтенсивності вираження певної оцінки. П'ятий етап безпосередня процедура підрахунку з використанням спеціалізованого програмного забезпечення або експертного кодування. Завершальний шостий етап інтерпретація отриманих результатів у контексті дослідницьких питань та теоретичних рамок.

2.3. Застосування контент-аналізу до соціальних мереж

Соціальні мережі як об'єкт контент-аналізу мають специфічні характеристики, що впливають на методологію дослідження. По-перше, обсяг даних у соціальних мережах набагато перевищує традиційні джерела, що робить ручну обробку практично неможливою та вимагає автоматизації [26]. По-друге, динамічний характер контенту соціальних мереж потребує майже реального часу обробки для виявлення актуальних тенденцій. По-третє, неформальний стиль комунікації, наявність жаргону, емодзі та специфічних скорочень ускладнює автоматичний аналіз.

Сучасні підходи до контент-аналізу соціальних мереж поєднують традиційні методи з технологіями машинного навчання та обробки природної мови [27]. Автоматична класифікація повідомлень за темами, аналіз тональності висловлювань, виявлення впливових користувачів та прогнозування поширення інформації є типовими задачами. Важливою складовою стає візуалізація результатів аналізу у формі графів соціальних зв'язків, хмар тегів, часових діаграм активності тощо. Етичні аспекти аналізу даних соціальних мереж включають питання конфіденційності користувачів та правомірності використання публічно доступної інформації у дослідницьких цілях.

3. Семантичний аналіз тексту

3.1. Поняття та рівні семантичного аналізу

Семантичний аналіз тексту є ключовим компонентом систем розуміння природної мови, спрямованим на виявлення змісту та значення текстових фрагментів [28]. На відміну від синтаксичного аналізу, який встановлює граматичну структуру речення, семантичний аналіз визначає смислові зв'язки між словами та інтерпретує текст у контексті предметної області. Семантика мови включає лексичну семантику, що визначає значення окремих слів, та композиційну семантику, яка пояснює, як значення складних виразів формується з значень їх частин.

У обробці природної мови виділяють кілька рівнів семантичного аналізу залежно від глибини інтерпретації [29]. Поверхневий семантичний аналіз ідентифікує основні сем античні ролі учасників ситуації, описаної у реченні, такі як агент дії, пацієнс, інструмент, місце та час. Глибинний семантичний аналіз будує формальне представлення змісту тексту у вигляді логічних форм або семантичних мереж, що дозволяє здійснювати логічний висновок та перевірку несуперечності. Прагматичний аналіз враховує контекст комунікації та наміри автора тексту, інтерпретуючи іронію, метафори та інші непрямі способи вираження думки.

3.2. Методи автоматичного семантичного аналізу

Автоматичний семантичний аналіз тексту базується на комбінації лінгвістичних знань та статистичних методів обробки даних. Традиційні підходи використовують семантичні словники та онтології, які формалізують знання про предметну область у вигляді концептів та відношень між ними [30]. Відомі ресурси такого типу включають

WordNet для англійської мови, який організує слова у синонімічні ряди та встановлює семантичні відношення між ними. Для морфологічно складних мов, включаючи українську, створення повноцінних семантичних ресурсів залишається актуальною задачею.

Статистичні методи семантичного аналізу базуються на гіпотезі дистрибутивності, згідно з якою слова зі схожими значеннями зустрічаються у схожих контекстах [31]. Векторні представлення слів, отримані за допомогою моделей Word2Vec, GloVe або FastText, кодують семантичну близькість у вигляді косинусної відстані між векторами. Такі представлення дозволяють здійснювати семантичний пошук, кластеризацію текстів за темами та виявлення семантичних аналогій. Нейромережеві моделі на основі трансформерів, зокрема BERT та його варіанти, досягають найвищої якості у задачах семантичного аналізу завдяки здатності враховувати контекст слова у реченні.

3.3. Аналіз тональності та виявлення емоцій

Аналіз тональності є окремим напрямком семантичного аналізу, спрямованим на визначення емоційного забарвлення тексту [32]. У найпростішому випадку тональність класифікується на позитивну, негативну та нейтральну, але існують більш деталізовані шкали, що враховують інтенсивність емоцій. Аналіз тональності широко застосовується у маркетингу для моніторингу відгуків про продукти та бренди, у політичній аналітиці для оцінки суспільних настроїв, та у фінансах для прогнозування коливань ринків на основі новинних потоків.

Методи аналізу тональності поділяються на підходи на основі словників та машинного навчання [33]. Словникові методи використовують заздалегідь підготовлені списки слів з позитивною та негативною конотацією, підраховуючи їх частоту у тексті. Проте такі методи не враховують контекст та граматичні конструкції, що можуть змінювати полярність висловлювання, наприклад заперечення або умовні конструкції. Методи машинного навчання тренуються на розмічених корпусах текстів і здатні виявляти складні патерни вираження тональності. Глибокі нейронні мережі демонструють найкращі результати, особливо у поєднанні з механізмами уваги, що дозволяють фокусуватися на найбільш значущих фрагментах тексту.

4. Оптимізація моделей даних для семантичного аналізу

4.1. Вимоги до організації даних

Ефективність систем семантичного аналізу контенту соціальних мереж критично залежить від способу організації вхідних даних та проміжних результатів обробки. Основними вимогами є швидкість доступу до даних, масштабованість системи при зростанні обсягів інформації, гнучкість схеми даних для адаптації до нових типів контенту та підтримка складних запитів для багатоаспектного аналізу [34]. Класичні реляційні бази даних забезпечують транзакційну цілісність та підтримку SQL запитів, але можуть не справлятися з навантаженням при обробці мільйонів повідомлень у реальному часі.

Для зберігання сирих даних соціальних мереж доцільно використовувати документо-орієнтовані бази даних, які дозволяють зберігати повідомлення у нативному JSON форматі без необхідності визначення жорсткої схеми [35]. Це забезпечує гнучкість при роботі з різнорідними даними з різних соціальних платформ, кожна з

яких має специфічну структуру метаданих. Для зберігання результатів семантичного аналізу, таких як виявлені сутності, тональність та теми, може застосовуватися реляційна модель, що полегшує агрегацію та аналітичні запити. Графові бази даних оптимальні для представлення соціального графа та зв'язків між користувачами, темами та контентом.

4.2. Індексування та пошук

Швидкий пошук за змістом повідомлень є фундаментальною вимогою систем аналізу контенту. Традиційні індекси баз даних на основі B-дерев ефективні для точного пошуку за ключами, але не підходять для повнотекстового пошуку з урахуванням морфології та синонімії [36]. Спеціалізовані пошукові системи, такі як Elasticsearch або Apache Solr, будують інвертовані індекси, що відображають кожне слово на список документів, у яких воно зустрічається, з додатковою інформацією про позицію та частоту.

Для семантичного пошуку застосовуються векторні індекси, що зберігають щільні представлення документів у багатовимірному просторі [37]. Пошук найближчих сусідів у векторному просторі дозволяє знаходити семантично подібні тексти навіть при відсутності спільних ключових слів. Апроксимаційні алгоритми пошуку найближчих сусідів, такі як HNSW або FAISS, забезпечують субліній ну складність запитів при прийнятній точності результатів. Гібридні підходи поєднують лексичний та семантичний пошук, ранжуючи результати з урахуванням обох факторів.

4.3. Паралелізм та розподілена обробка

Обробка великих обсягів текстових даних вимагає паралелізації обчислень та розподіленої архітектури системи. Фреймворки розподіленої обробки даних, такі як Apache Spark або Apache Flink, дозволяють виконувати семантичний аналіз на кластерах серверів з автоматичним балансуванням навантаження та відмовостійкістю [38]. Дані розбиваються на партиції, кожна з яких обробляється незалежно, після чого результати агрегуються. Така архітектура забезпечує лінійну масштабованість продуктивності при додаванні нових вузлів до кластеру.

Для задач потокової обробки, коли необхідно аналізувати дані у реальному часі по мірі їх надходження зі соціальних мереж, застосовуються системи потокової обробки [39]. Вони підтримують віконні операції для обчислення агрегатів за часовими інтервалами та стейтфул обробку з можливістю накопичення проміжних результатів. Критичним аспектом є забезпечення семантики обробки exactly-once для запобігання дублюванню або втраті повідомлень. Інтеграція потокової та пакетної обробки у lambda-архітектурі дозволяє поєднувати швидкість реагування з точністю результатів.

Висновки

У роботі проведено систематичний аналіз моделей організації даних та їх застосування для задач семантичного аналізу контенту соціальних мереж. Встановлено, що вибір моделі даних суттєво впливає на ефективність системи та повинен враховувати специфіку предметної області. Класичні ієрархічна та мережева моделі забезпечують високу швидкодію, але недостатньо гнучкі для роботи зі складними та динамічними структурами соціальних даних. Реляційна модель пропонує

універсальність та потужні засоби запитів, але може мати обмеження масштабованості при дуже великих обсягах даних.

Сучасні NoSQL рішення, зокрема документо-орієнтовані та графові бази даних, краще відповідають вимогам обробки контенту соціальних мереж завдяки гнучкості схеми та оптимізації для розподілених систем. Гібридні архітектури, що поєднують різні типи сховищ залежно від характеру даних та операцій, є найбільш перспективним напрямком. Для ефективного семантичного аналізу критичне значення має організація індексів, що забезпечують швидкий пошук як за лексичними, так і за семантичними ознаками.

Контент-аналіз та семантичний аналіз тексту є потужними інструментами вивчення соціальних комунікацій, але їх практичне застосування вимагає значних обчислювальних ресурсів та ретельного проектування архітектури системи. Подальші дослідження доцільно спрямувати на розробку спеціалізованих моделей даних, оптимізованих для конкретних задач семантичного аналізу, та на інтеграцію нових методів машинного навчання з традиційними підходами організації даних.

Список використаної літератури:

1. Manyika J. Big data: The next frontier for innovation, competition, and productivity / J. Manyika, M. Chui, B. Brown // McKinsey Global Institute. – 2011. – 156 p.
2. Стадник А. В. Аналіз великих даних соціальних мереж / А. В. Стадник, О. І. Пеньковський // Вісник Національного університету “Львівська політехніка”. Інформаційні системи та мережі. – 2016. – № 854. – С. 357–367.
3. Connolly T. Database Systems: A Practical Approach to Design, Implementation and Management / T. Connolly, C. Begg. – 6th ed. – Pearson, 2015. – 1440 p.
4. Neuendorf K. A. The Content Analysis Guidebook / K. A. Neuendorf. – 2nd ed. – SAGE Publications, 2017. – 472 p.
5. Pang B. Opinion mining and sentiment analysis / B. Pang, L. Lee // Foundations and Trends in Information Retrieval. – 2008. – Vol. 2, № 1-2. – P. 1–135.
6. Date C. J. An Introduction to Database Systems / C. J. Date. – 8th ed. – Pearson, 2004. – 1024 p.
7. Elmasri R. Fundamentals of Database Systems / R. Elmasri, S. B. Navathe. – 7th ed. – Pearson, 2016. – 1272 p.
8. Garcia-Molina H. Database Systems: The Complete Book / H. Garcia-Molina, J. D. Ullman, J. Widom. – 2nd ed. – Pearson, 2009. – 1248 p.
9. Tsichritzis D. C. The ANSI/X3/SPARC DBMS Framework / D. C. Tsichritzis, A. Klug // Information Systems. – 1978. – Vol. 3, № 3. – P. 173–191.
10. Silberschatz A. Database System Concepts / A. Silberschatz, H. F. Korth, S. Sudarshan. – 7th ed. – McGraw-Hill, 2020. – 1376 p.
11. Ramakrishnan R. Database Management Systems / R. Ramakrishnan, J. Gehrke. – 3rd ed. – McGraw-Hill, 2003. – 1065 p.
12. Bachman C. W. The programmer as navigator / C. W. Bachman // Communications of the ACM. – 1973. – Vol. 16, № 11. – P. 653–658.
13. Taylor R. W. A comparison of the CODASYL and relational approaches to data-base management / R. W. Taylor, R. L. Frank // Proceedings of the 1976 Conference on Data: Abstraction, Definition and Structure. – ACM, 1976. – P. 52–67.
14. Ozsoyoglu G. A summary of the redesign of the CODASYL sets / G. Ozsoyoglu, H. Wang // Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. – 1993. – P. 503–504.
15. Codd E. F. A relational model of data for large shared data banks / E. F. Codd // Communications of the ACM. – 1970. – Vol. 13, № 6. – P. 377–387.
16. Maier D. The Theory of Relational Databases / D. Maier. – Computer Science Press, 1983. – 637 p.
17. Stonebraker M. The design of POSTGRES / M. Stonebraker, L. A. Rowe // Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data. – 1986. – P. 340–355.
18. Atkinson M. The object-oriented database system manifesto / M. Atkinson, F. Bancilhon, D. DeWitt // Proceedings of the First International Conference on Deductive and Object-Oriented Databases. – 1989. – P. 223–240.

19. Cattell R. Scalable SQL and NoSQL data stores / R. Cattell // ACM SIGMOD Record. – 2011. – Vol. 39, № 4. – P. 12–27.
20. Robinson I. Graph Databases: New Opportunities for Connected Data / I. Robinson, J. Webber, E. Eifrem. – 2nd ed. – O'Reilly Media, 2015. – 238 p.
21. Krippendorff K. Content Analysis: An Introduction to Its Methodology / K. Krippendorff. – 4th ed. – SAGE Publications, 2019. – 472 p.
22. Berelson B. Content Analysis in Communication Research / B. Berelson. – Free Press, 1952. – 220 p.
23. Holsti O. R. Content Analysis for the Social Sciences and Humanities / O. R. Holsti. – Addison-Wesley, 1969. – 235 p.
24. Riffe D. Analyzing Media Messages: Using Quantitative Content Analysis in Research / D. Riffe, S. Lacy, F. Fico. – 3rd ed. – Routledge, 2014. – 280 p.
25. Манаєв О. Т. Контент-аналіз матеріалів засобів масової інформації / О. Т. Манаєв. – К.: Центр вільної преси, 1998. – 200 с.
26. Stieglitz S. Social media analytics / S. Stieglitz, M. Dang-Xuan, A. Bruns // Business & Information Systems Engineering. – 2014. – Vol. 6, № 2. – P. 89–96.
27. Nguyen T. H. Social media analytics for enterprises / T. H. Nguyen, K. Shirai, J. Velcin // ACM Computing Surveys. – 2015. – Vol. 48, № 1. – P. 1–37.
28. Jurafsky D. Speech and Language Processing / D. Jurafsky, J. H. Martin. – 3rd ed. – Pearson, 2021. – 600 p.
29. Allen J. Natural Language Understanding / J. Allen. – 2nd ed. – Benjamin Cummings, 1995. – 654 p.
30. Navigli R. Word sense disambiguation: A survey / R. Navigli // ACM Computing Surveys. – 2009. – Vol. 41, № 2. – P. 1–69.
31. Turney P. D. From frequency to meaning: Vector space models of semantics / P. D. Turney, P. Pantel // Journal of Artificial Intelligence Research. – 2010. – Vol. 37. – P. 141–188.
32. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions / B. Liu. – 2nd ed. – Cambridge University Press, 2020. – 492 p.
33. Mohammad S. M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text / S. M. Mohammad // Emotion Measurement. – Elsevier, 2016. – P. 201–237.
34. Stonebraker M. What goes around comes around / M. Stonebraker, J. M. Hellerstein // Readings in Database Systems. – 4th ed. – MIT Press, 2005. – P. 2–41.
35. Strauch C. NoSQL Databases / C. Strauch, U. L. S. Sites, W. Kriha. – Stuttgart Media University, 2011. – 149 p.
36. Zobel J. Inverted files for text search engines / J. Zobel, A. Moffat // ACM Computing Surveys. – 2006. – Vol. 38, № 2. – P. 1–56.
37. Johnson J. Billion-scale similarity search with GPUs / J. Johnson, M. Douze, H. Jégou // IEEE Transactions on Big Data. – 2021. – Vol. 7, № 3. – P. 535–547.
38. Zaharia M. Apache Spark: A unified engine for big data processing / M. Zaharia, R. S. Xin, P. Wendell // Communications of the ACM. – 2016. – Vol. 59, № 11. – P. 56–65.
39. Carbone P. Apache Flink: Stream and batch processing in a single engine / P. Carbone, A. Katsifodimos, S. Ewen // Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. – 2015. – Vol. 38, № 4. – P. 28–38.

References:

1. Manyika J., Chui M., Brown B. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.
2. Stadnyk A. V., Penkovskyi O. I. (2016). Analysis of big data in social networks. *Bulletin of Lviv Polytechnic National University. Information Systems and Networks,* 854, 357–367 [in Ukrainian].
3. Connolly T., Begg C. (2015), Database Systems: A Practical Approach to Design, Implementation and Management, 6th ed. Pearson.
4. Neuendorf K. A. (2017). The Content Analysis Guidebook, 2nd ed. AGE Publications.
5. Pang B., Lee L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1–135.
6. Date C. J. (2004). An Introduction to Database Systems, 8th ed. Pearson.
7. Elmasri R., Navathe S. B. (2016). Fundamentals of Database Systems, 7th ed. Pearson.
8. Garcia-Molina H., Garcia-Molina H., Ullman J. D., Widom J. (2009). Database Systems: The Complete Book, 2nd ed. Pearson.
9. Tsichritzis D. C., Klug A. (1978). The ANSI/X3/SPARC DBMS Framework. *Information Systems*, 3(3), 173–191.
10. Silberschatz A., Korth H. F., Sudarshan S. (2020). Database System Concepts, 7th ed. McGraw-Hill.

11. Ramakrishnan R., Gehrke J. (2003). Database Management Systems, 3rd ed. McGraw-Hill.
12. Bachman C. W. (1973). The programmer as navigator. *Communications of the ACM*, 16, 11, 653–658.
13. Taylor R. W., FrankR. L. (1976) A comparison of the CODASYL and relational approaches to data-base management. *Proceedings of the 1976 Conference on Data: Abstraction, Definition and Structure*, ACM, 52–67.
14. Ozsoyoglu G., Wang H. (1993). A summary of the redesign of the CODASYL sets. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 503–504.
15. Codd E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377–387.
16. Maier D. (1983). The Theory of Relational Databases. Computer Science Press.
17. Stonebraker M., Rowe L. A. (1986). The design of POSTGRES. *Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data*, 340–355.
18. Atkinson M., Bancilhon F., De Witt. (1989). The object-oriented database system manifesto. *Proceedings of the First International Conference on Deductive and Object-Oriented Databases,* 223–240.
19. Cattell R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39(4), 12–27.
20. Robinson I., Webber J., Eifrem E. (2015) Graph Databases: New Opportunities for Connected Data, 2nd ed. O'Reilly Media.
21. Krippendorff K. (2019) Content Analysis: An Introduction to Its Methodology, 4th ed. SAGE Publications.
22. Berelson B. (1952). Content Analysis in Communication Research. Free Press.
23. Holsti O. R. (1969). Content Analysis for the Social Sciences and Humanities. Addison-Wesley.
24. Riffe D., Lacy S., Fico F. (2014). Analyzing Media Messages: Using Quantitative Content Analysis in Research, 3rd ed. Routledge.
25. Manaiev O. T. (1998). Content Analysis of Mass Media Materials. Kyiv: Free Press Center. [in Ukrainian].
26. Stieglitz S., Dang-Xuan M., Bruns A. (2014). Social media analytics. *Business & Information Systems Engineering*, 6(2), 89–96.
27. Nguyen T. H., Shirai K., Velcin J. (2015). Social media analytics for enterprises. *ACM Computing Surveys,* 48(1), 1–37.
28. Jurafsky D., Martin J. H. (2021). Speech and Language Processing, 3rd ed. Pearson.
29. Allen J. (1995). Natural Language Understanding, 2nd ed. Benjamin Cummings.
30. Navigli R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2), 1–69.
31. Turney P. D. Pantel P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.
32. Liu B. (2020). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 2nd ed. Cambridge University Press
33. Mohammad S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. Emotion Measurement. Elsevier, 201–237.
34. Stonebraker M., Hellerstein J. M. (2005). What goes around comes around. Readings in Database Systems, 4th ed. MIT Press, 2–41.
35. Strauch C., Sites U. L. S., Kriha W. (2011). NoSQL Databases. Stuttgart Media University.
36. Zobel J. Inverted files for text search engines / J. Zobel, A. Moffat // ACM Computing Surveys. – 2006. – Vol. 38, № 2. – P. 1–56.
37. Johnson J., Douze M., Jégou H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.
38. Zaharia M., Xin R. S., Wendell P. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11), 56–65.
39. Carbone P., Katsifodimos A., Ewen S. (2015). Apache Flink: Stream and batch processing in a single engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 38(4), 28–38.

KUTSENKO Oleksandr,
Postgraduate, Department of Informatics and Applied Mathematics, The Bohdan Khmelnytsky National University of Cherkasy, Ukraine
PISKUN Oleksandr,
Candidate of Technical Sciences, Associate Professor, Head of Department of Applied Mathematics and Informatics, Bohdan Khmelnytsky National University of Cherkasy

DATA ORGANIZATION MODELS AND THEIR OPTIMIZATION FOR SEMANTIC ANALYSIS OF SOCIAL MEDIA CONTENT

Abstract. Introduction. This paper investigates fundamental approaches to data organization in information systems and their application for optimizing semantic analysis of social media content. Classical data models – hierarchical, network, and relational – are analyzed, and their advantages and limitations in the context of processing large volumes of textual information are identified. The methodology of content analysis as a tool for studying social communications and semantic text analysis as a stage of automatic natural language understanding are examined. The relationship between the choice of data organization model and the efficiency of semantic processing algorithms for social media content is established.

The exponential growth of digital information, particularly in social networks, creates both new opportunities and challenges for data analysis. Effective processing of such data arrays requires not only powerful computing resources but also optimal data organization. The choice of data organization model directly affects the speed of information access, the complexity of processing algorithms, and the quality of final analysis results. Meanwhile, there is a growing need for automated content analysis systems capable of identifying hidden patterns, trends, and semantic connections in large arrays of textual data.

Content analysis as a research method originated in sociology and journalism but has transformed into a powerful tool for automatic text processing with the development of computing technology. Semantic analysis, which is a component of content analysis, allows identifying not only explicit but also hidden content of messages, which is especially important for understanding public sentiments, identifying trends, and predicting user behavior in social networks.

Purpose. The aim of this article is to analyze classical data organization models and determine their optimality for the tasks of semantic analysis of social network content.

Results. The hierarchical model represents data in a tree-like structure with strict subordination relationships but has limited flexibility for representing complex many-to-many connections. The network model extends the hierarchical approach by allowing multiple parent relationships but increases system complexity. The relational model, based on set theory and relational algebra, provides data independence and declarative query language but may have performance limitations for certain operations.

Modern data models, including document-oriented and graph databases, better meet the requirements of social media content processing due to schema flexibility and optimization for distributed systems. Hybrid architectures combining different types of storage depending on data nature and operations are identified as the most promising direction. For effective semantic analysis, the organization of indexes that provide fast search by both lexical and semantic features is critically important.

Conclusion. The paper examines content analysis stages including source selection, sampling, unit of analysis identification, and results interpretation. Semantic analysis methods are reviewed, including sentiment analysis, entity recognition, and topic modeling. The relationship between data organization model choice and semantic analysis algorithm efficiency is established. Recommendations for optimizing data structures for processing large volumes of social media content are provided.

The research demonstrates that no single data model is universally optimal for all aspects of social media content analysis. Document-oriented databases are suitable for storing raw heterogeneous data, relational databases for storing structured analysis results, and graph databases for representing social connections. Distributed processing and stream processing technologies are essential for real-time analysis of large-scale social media data.

Keywords: data models, data organization, content analysis, semantic text analysis, social networks, optimization, hierarchical model, network model, relational model, NoSQL databases.

Одержано редакцією 28.10.2022 р.
Прийнято до публікації 17.11.2022 р.