ДОДАТОК А

СХЕМА АЛГОРИТМУ ПРОГНОЗУВАННЯ ВЕБ-СТОРІНОК

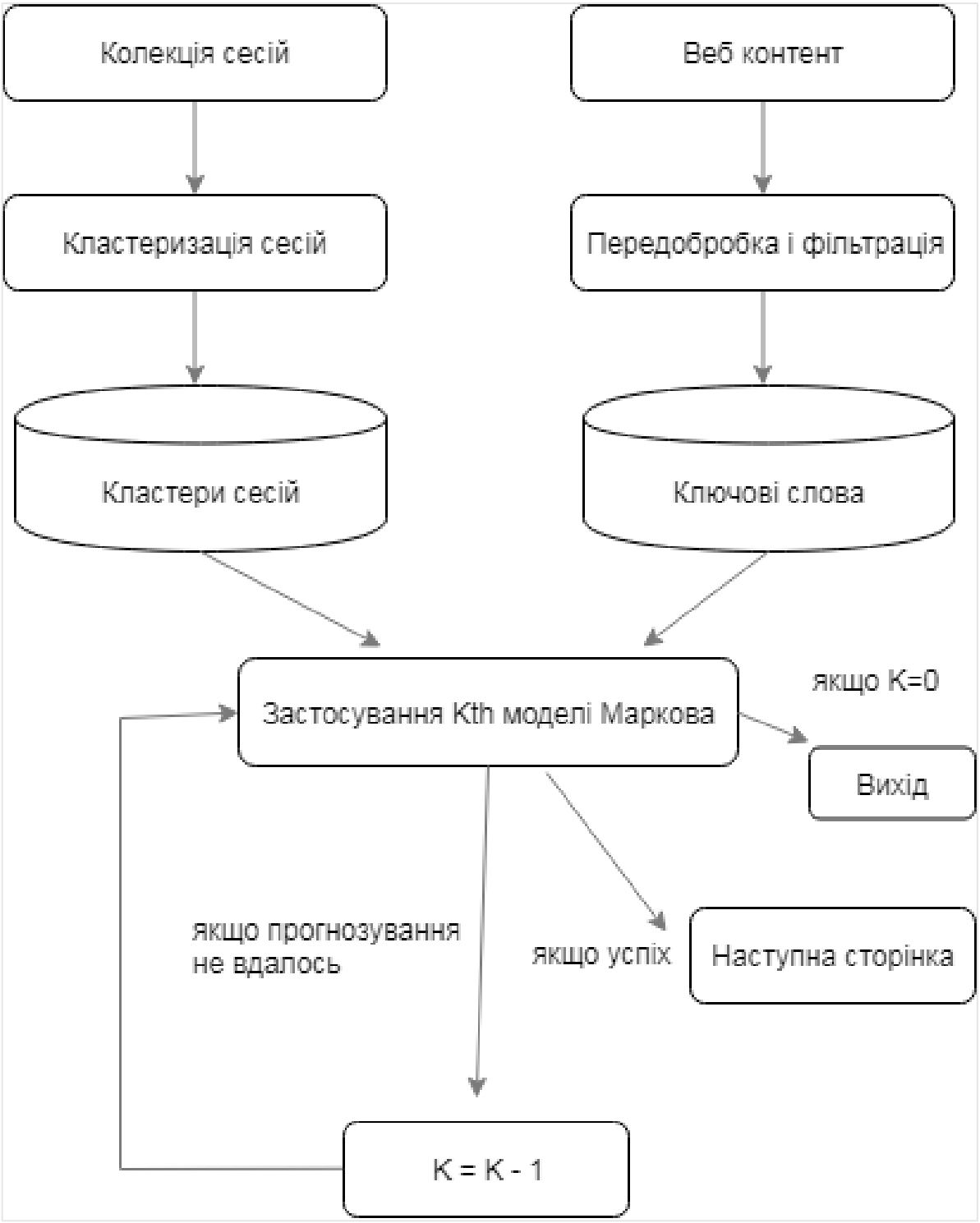


Рисунок А.1 – Схема алгоритму прогнозування

Алгоритм прогнозування веб-сторінок:

− сесія веб-користувача збирається у вигляді вектора сеансу з веб-сайту;

− виконується кластеризація сесії ієрархічним підходом;

− введення особливості – набір ключових слів для кожної сторінки веб-контенту. Збирається зміст кожної веб-сторінки на предмет ключових слів;

− попередня обробка і фільтрація отриманого контенту;

− встановлюємо верхню межу і нижню межу частоти слів, які з'являються в змісті. Таким чином вектор має колекцію ключових слів по сторінках;

− функція колекції ключових слів, які отримуються з попередньої сторінки сеансу порівнює ключові слова сторінок у вектор. Найбільш подібна сторінка буде наступною цільовою сторінкою сесії. Ця сторінка повертається до функції;

− генеруємо всі порядки марковських моделей і використовуємо їх колективно в прогнозі;

− якщо прогнозування з використанням моделі Маркова вищого порядку не вдається, то марківська модель розглядається з використанням нового сеансу довжини K–1. Тобто зниженням порядку моделі;

− процес повторюється до досягнення марківської моделі першого порядку або коли прогнозування не відбувається.

ДОДАТОК Б

СЛАЙДИ ПРЕЗЕНТАЦІЇ

1

Міністерство освіти і науки України
Харківський національний університет
радіоелектроніки

Атестаційна робота магістра

Дослідження методів аналізу даних поведінки користувачів веб орієнтованих систем

Науковий керівник:
к.т.н, доц. каф. ПІ Назаров О.С.
Виконав:
студент групи ІПЗ мзд-17-1 Іванов О.В.

2019

2

Актуальність теми

- використання в електронній комерції, оптимізація процесу онлайн-покупок та збільшення продажів;
- системи контекстної реклами;
- поліпшення змісту і структури веб сайтів;
- виявлення внутрішніх загроз і цільових атак;
- запобігання фінансового шахрайства;
- використання в системах веб-аналітики.

3

Мета дослідження - виявлення особливостей та вдосконалення існуючих методів аналізу даних поведінки веб-користувачів.

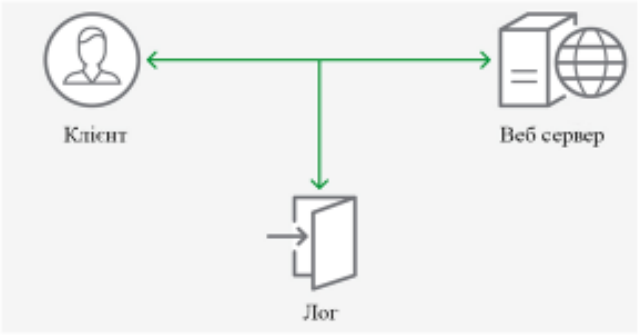
Задачами дослідження є:

- виявлення шляхів та засобів удосконалення існуючих та запропонування нових методів аналізу даних поведінки веб-користувачів;
- удосконалення методу прогнозування поведінки веб-користувачів з використанням моделей Маркова;
- оцінка ефективності запропонованого методу шляхом проведення аналітичного та експериментального дослідження.

4

Об'єктом дослідження є серверні логі, процес обробки та методи їх аналізу.

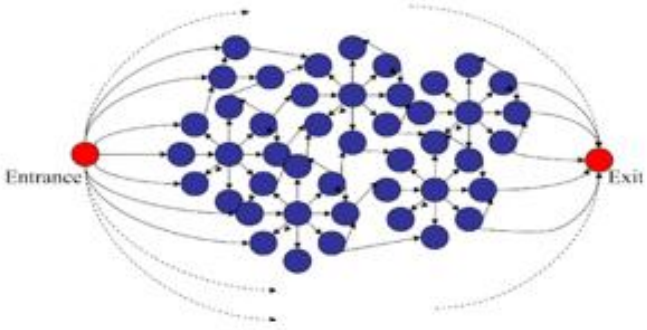
Предметом дослідження є моделі, методи і алгоритми для виявлення і визначення поведінки веб-користувачів на основі серверних логів.



5

Репрезентативні змінні

- структура;
- контент;
- сеанс користувача.



6

Основні джерела даних

Журнали веб-сервера

143.158.26.50 - - [01/Aug/1995:00:18:00 -0400] "GET /shuttle/missions/61-b/mission-61-b.html HTTP/1.0" 200 6468

blv-pm11-ip17.halcyon.com - - [01/Aug/1995:00:20:01 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786

7

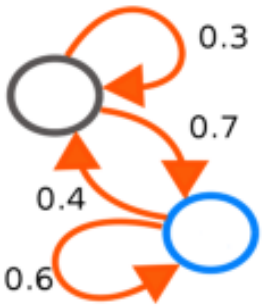
Попередня обробка даних

- видалення глобальних і локальних шумів;
- записи з кодом стану HTTP понад 299 або менше 200 видаляються;
- видалення записів, що не відображають активність користувача;
- ідентифікація сеансу користувача.

8

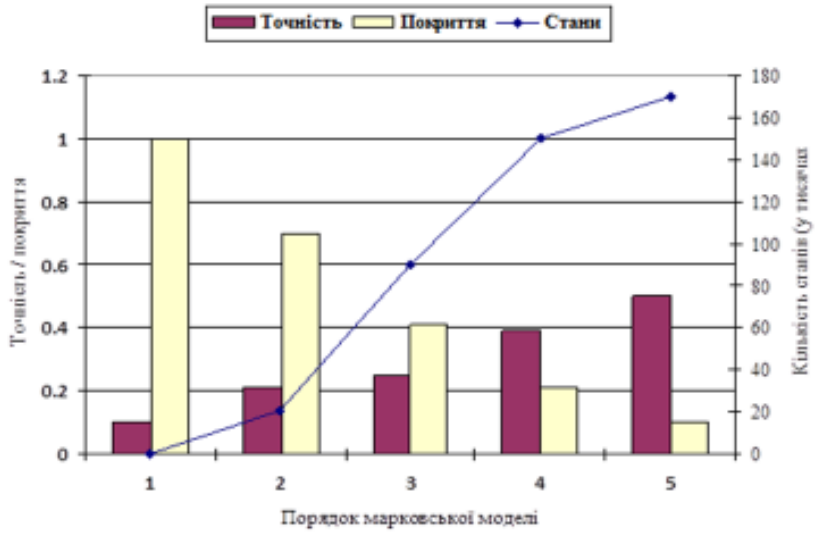
Ланцюги Маркова

Марковський процес - випадковий процес, поведінка якого залежить тільки від поточного стану, тобто не залежить від минулого



9

Графік порівняння точності, покриття та розміру моделі з порядком марковської моделі



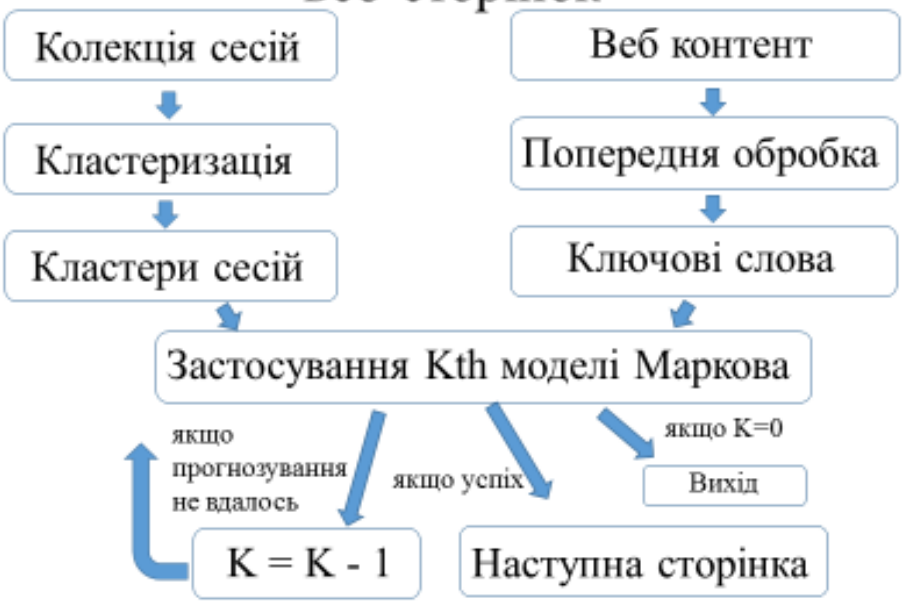
10

Запропонований підхід

- сесія веб-користувача збирається у вигляді вектора сеансу з веб-сайту;
- виконується кластеризація сесії ієрархічним підходом;
- введення особливості - набір ключових слів для кожної сторінки веб-контенту;
- генеруємо всі порядки марковських моделей і використовуємо їх колективно в прогнозі.

11

Схема алгоритму прогнозування веб-сторінок



12

Резюме набору даних

Було розглянуто два набори даних - NASA та штучний.

	NASA	Штучні
Всього сесій	50,000	20,000
Середня тривалість сесії	6.4	4.5
Кількість сторінок	2266	16
Час датасета	серпень 1995	свій

13

Результати експерименту

Для різних навчальних наборів спостерігається підвищення тренувальних даних, а також підвищення їх точності, але на дуже малу частку.

Так як після 50% даних точність змінюється незначно в різних значеннях морківського порядку.

14

Усі значення K-порядку для 60% розміру набору даних

	Штучний набір даних		NASA DataSet
Порядок	Журнали	Журнали + вміст	Журнали
Перший	0.062	0.085	0.91
Другий	0.1213	0.132	0.088
Третій	0.564	0.543	0.014
Четвертий	0.4102	0.208	0.138
Точність	0.1953	0.2050	0.086

15

Висновки

- марковська модель є найбільш часто використовуваною моделлю прогнозування через її високу точність;
- низький порядок марковських моделей має більш високу точність і менше покриття;
- моделі вищого порядку мають ряд обмежень, пов'язаних з вищою складністю стану, зменшеним охопленням, іноді навіть гіршою точністю прогнозування.

16

Апробація результатів роботи

- Дослідження основних джерел, процесів і методів обробки даних поведінки користувачів веб орієнтованих систем. Новини науки: дослідження, наукові відкриття, високі технології: зб. наук. праць «ΛΌГОΣ» з матеріалами міжнар. наук.-практ. конф., м. Харків, 31 березня, 2019 р. Харків : ГО «Європейська наукова платформа». ISBN 978-617-7171-80-4. С. 15-17;
- Дослідження методів аналізу даних поведінки користувачів веб орієнтованих систем. Збірник наукових робіт «ADVANCED OF SCIENCE» (Карлові Вари, Чехія) з матеріалами міжнар. наук.-практ. конф. «Discovery Science», м. Київ, м. Карлові Вари, 5 квітня 2019 р.. ISBN 978-80-7534-078-8. С. 213-218;

17

Апробація результатів роботи (продовження)

- Прогнозування поведінки користувачів веб орієнтованих систем на основі аналізу серверних логів. Міжнародна наукова інтернет-конференція "Інформаційне суспільство: технологічні, економічні та технічні аспекти становлення (випуск 37)" / Збірник тез доповідей: випуск 37 (м. Тернопіль, 2 квітня 2019 р.). – Тернопіль. – 2019. ISSN 2522-932X. С. 23-25;
- Попередня обробка даних для аналізу поведінки користувачів веб орієнтованих систем. Прикладні наукові розробки та теоретичні дослідження XXI століття: зб. наук. праць «ΛΌГОΣ» з матеріалами міжнар. наук.-практ. конф., м. Вінниця, 15 квітня, 2019 р. Вінниця: ГО «Європейська наукова платформа», 2019. ISBN 978-617-7171-80-4. С 72-74.

18

Дякую за увагу

ДОДАТОК В

АПРОБАЦІЯ РЕЗУЛЬТАТІВ РОБОТИ

Іванов О.В. Дослідження основних джерел, процесів і методів обробки даних поведінки користувачів веб-орієнтованих систем. Новини науки: дослідження, наукові відкриття, високі технології: зб. наук. праць «ΛΌГОΣ» з матеріалами міжнар. наук.-практ. конф., м. Харків, 31 березня, 2019 р. Харків : ГО «Європейська наукова платформа». ISBN 978-617-7171-80-4. С. 15-17

В даний час зростання World Wide Web перевищує всі очікування. Інтернет зростає з кожним днем, а разом з тим зростає і активність відвідувачів, які проводять досить багато часу в ньому. Можливість отримання цінної інформації від такого великого об'єму даних вимагає актуальних підходів і методів аналізу.

Інтелектуальний аналіз даних є процесом вилучення неявної інформації та виявлення змістовних шаблонів, профілів і тенденцій з великих баз даних. Цей ітеративний процес виявляється цінною стратегією для розуміння активності користувачів в інтернеті.

Головним чином, існують чотири типи джерел даних, в яких дані про використання записуються на різних рівнях: рівень клієнта, рівень браузера, збір на рівні сервера та проксі.

Збір клієнтського рівня: На цьому рівні дані збираються разом за допомогою скриптів. Ці дані показують поведінку одного користувача на одному сайті. Збір даних на стороні клієнта вимагає участі користувача для ввімкнення скриптів або аплетів. Перевагою збору даних на стороні клієнта є те, що цей збір може захопити всі кліки, включаючи натискання кнопки назад або перезавантаження сторінки.

Колекція рівня браузера: Другий спосіб збору даних – це модифікація браузера. Він показує поведінку одного користувача на декількох сайтах. Можливості збору даних розширюються шляхом зміни вихідного коду існуючого браузера. Вони надають набагато більш різноманітні дані, оскільки розглядають поведінку одного користувача на декількох сайтах.

Збірка рівня сервера: Журнал веб-сервера [1] зберігає поведінку декількох користувачів по одному сайту. Ці файли журналів можуть зберігатися в загальному форматі журналу або розширеному форматі журналу. Журнали сервера не можуть зберігати кешовані перегляди сторінок. Іншим методом, що використовується для збору даних на рівні сервера, є перехоплення пакетів TCP / IP. Пакетні переглядачі працюють шляхом моніторингу трафіку мережевої роботи і безпосереднього отримання даних про використання.

Журнали веб-сервера – це текстові файли (ASCII) і незалежні від сервера. Існують деякі відмінності між серверним програмним забезпеченням, але традиційно існує кілька типів серверних журналів: журнал доступу та журнал помилок. Запис в ці журнали ведеться в певних форматах. Common Log Format (CLF) створено для відстеження запитів, які відбуваються на веб-сайті в хронологічному порядку. Він містить IP-адресу клієнта, ім'я хоста, ім'я користувача, позначку часу, ім'я файлу та розмір файлу.

Збір на рівні проксі: проксі-сервери використовуються провайдером послуг інтернету для надання клієнтам доступу до World Wide Web. Ці сервери зберігають поведінку декількох користувачів на декількох сайтах. Такі серверні функції, як кеш-сервер, можуть створювати кешовані перегляди сторінок. Прогнозуючи схему використання відвідувача інтелектуальний аналіз даних покращує якість послуг електронної комерції, персоналізує інтернет або підвищує продуктивність веб-структури та веб-сервера.

Дані, що знаходяться в лог-журналах, не можуть використовуватися для аналізу в тому вигляді в якому вони існують і повинні бути оброблені спеціальним чином.

Процес отримання даних про поведінку веб-користувачів можна розділити на три категорії. Перша – це попередня обробка, в процесі якої сесії веб-користувачів виводяться з джерел даних, якими є веб-журнали, які зберігають дії кожного відвідувача сайту. Такі файли можуть містити мільйони записів залежно від трафіку веб-сайту це і є головне джерело даних про поведінку людини.

Друга категорія це пошук шаблонів в даних. Цей процес здійснюється за допомогою стандартних методів інтелектуального аналізу даних, таких як пошук асоціативних правил або послідовних шаблонів [2].

На третьому етапі інформаційні фільтри які базуються на знаннях домену і структурах веб-сайту застосовуються до шаблонів аналізу в пошуках цікавих шаблонів.

Фаза попередньої обробки даних виконується з використанням перетвореного файлу журналу, який був очищений шляхом видалення всіх непотрібних, нерегулярних і відсутніх даних з оригінального загального файлу журналу. Після початкової попередньої обробки фільтр веб-сеансу використовують до перетвореного файлу журналу для вилучення ознак. Метою фільтра є агрегування всіх запитів користувачів у сесії в єдиний набір змінних.

Для виявлення цікавих шаблонів застосовуються статистичні методи, а також методи інтелектуального аналізу даних – аналіз шляхів, правило асоціації, послідовні структури та правила кластерів та класифікації.

Фаза видобування даних включає дві підфази: дескриптивний аналіз і аналіз за допомогою штучного інтелекту. Використовується описовий аналіз підбивання підсумків, методів кластеризації та асоціативних правил для генерування набору даних, отримання уявлення про характеристики користувачів і описати основні шаблони поведінки користувачів. Аналіз за допомогою штучного інтелекту [3] використовується для прогнозування.

Фаза аналізу моделей включає інтерпретацію даних та оцінку результатів. Цей етап необхідний для визначення значущих результатів з результатів фази аналізу даних.

Список використаних джерел:

1. [Електронний ресурс] – Режим доступу до ресурсу: https://httpd.apache.org/docs/1.3/logs.html;

2. [Електронний ресурс] – Режим доступу до ресурсу: http://data-mining.philippe-fournier-viger.com/introduction-sequential-pattern-mining/;

3. [Електронний ресурс] – Режим доступу до ресурсу: https://www.i-scoop.eu/artificial-intelligence-cognitive-computing/

Іванов О.В. Дослідження методів аналізу даних поведінки користувачів веб-орієнтованих систем. Збірник наукових робіт «ADVANCED OF SCIENCE» (Карлові Вари, Чехія) з матеріалами міжнар. наук.-практ. конф. «Discovery Science», м. Київ, м. Карлові Вари, 5 квітня 2019 р.. ISBN 978-80-7534-078-8. С. 213-218.

З ранніх часів цивілізації людство зіткнулося з проблемою розуміння самого себе. Трейдери передбачають потреби людей, політики розраховують кроки з найкращим політичним результатом, а генерали вирішують позицію армії. Люди живуть разом у суспільствах, що складають складні системи взаємозалежності. Одним з етапів для побудови кращого суспільства є наявність достатніх знань про людську поведінку [1].

Поведінка перегляду веб-користувачів може бути описана трьома видами даних: веб-структурою, веб-контентом і веб-сеансом користувача. Перший безпосередньо пов'язаний з навколишнім середовищем. Третій описує потік кліків який виконує кожен веб-користувач під час свого відвідування веб-сайту.

Data Mining – це процес виявлення в "сирих" даних раніше невідомих нетривіальних практично корисних і доступних інтерпретації, необхідних для прийняття рішень в різних сферах людської діяльності. Data Mining є одним з кроків Knowledge Discovery in Databases.

Інформація, знайдена в процесі застосування методів Data Mining, повинна бути нетривіальною і раніше невідомої, наприклад, середні продажі не є такими.

Знання повинні описувати нові зв'язки між властивостями, передбачати значення одних ознак на основі інших і т.д. Знайдені знання повинні бути застосовні і на нових даних з деякою мірою вірогідності. Корисність полягає в тому, що ці знання можуть приносити певну вигоду при їх застосуванні. Знання повинні бути в зрозумілій для користувача не математика вигляді. Наприклад, найпростіше сприймаються людиною логічні конструкції "якщо ... то ...". Більш

того, такі правила можуть бути використані в різних СУБД у якості SQL-запитів. У разі, коли витягнуті знання непрозорі для користувача, повинні існувати методи обробки поста, що дозволяють привести їх до інтерпретованих.

Алгоритми, що використовуються в Data Mining, вимагають великої кількості обчислень. Раніше це було фактором що стримує широке практичне застосування Data Mining, проте сьогоднішнє зростання продуктивності сучасних процесорів зняв гостроту цієї проблеми. Тепер за прийнятний час можна провести якісний аналіз сотень тисяч і мільйонів записів.

Класифікація – це віднесення об'єктів (спостережень, подій) до одного з заздалегідь відомих класів.

Регресія, в тому числі завдання прогнозування. Встановлення залежності безперервних вихідних від вхідних змінних.

Кластеризація – це групування об'єктів (спостережень, подій) на основі даних (властивостей), що описують сутність цих об'єктів. Об'єкти усередині кластера повинні бути "схожими" один на одного і відрізнятися від об'єктів, які увійшли в інші кластери. Чим більше схожі об'єкти усередині кластера і чим більше відмінностей між кластерами, тим точніше кластеризація.

Асоціація – виявлення закономірностей між пов'язаними подіями. Прикладом такої закономірності служить правило, яке вказує, що з події X слід подія Y. Такі правила називаються асоціативними. Вперше ця задача була запропонована для знаходження типових шаблонів покупок, що здійснюються в супермаркетах, тому іноді її ще називають аналізом ринкової корзини (market basket analysis).

Послідовні шаблони – встановлення закономірностей між пов'язаними в часі подіями, тобто виявлення залежності, що якщо відбудеться подія X, то через заданий час відбудеться подія Y.

Аналіз відхилень – виявлення найбільш нехарактерних шаблонів.

Проблеми бізнес аналізу формулюються по-іншому, але рішення більшості з них зводиться до тієї чи іншої задачі Data Mining або до їх комбінації. Наприклад, оцінка ризиків це вирішення завдання регресії або класифікації, сегментація ринку

кластеризація, стимулювання попиту – асоціативні правила. Фактично, завдання Data Mining є елементами, з яких можна зібрати рішення переважної більшості реальних бізнес завдань. Data Mining має мультидисциплінарний характер.

Однак, Data Mining не є срібною кулею для багатьох проблем, включаючи аналіз поведінки людини. Процес ієрархічного каскаду інтелектуального аналізу даних [4] показав, що багато ітерацій керованих людиною необхідні для остаточного налаштування моделі на дані і застосування її в режимі прогнозування. В даний час нові комп'ютерні удосконалення дозволяють мати більш автоматичний процес налаштування моделей машинного навчання, що складають нове покоління інтелектуальних програм. Редакція журналу Wired під назвою "Кінець теорії: потік даних робить застарілим науковий метод" стверджує, що сучасні інтелектуальні додатки є досить потужними, щоб обробляти будь-яку реальну складність, роблячи багато теорій застарілими [5]. Очевидно, що вибуховий приріст даних вплинув на точність багатьох методів моделювання. Отже, ця пропозиція здається занадто оптимістичною, щоб бути правдою і є фундаментальні причини відкинути цю надзвичайно наївну пропозицію.

Незважаючи на те, що сучасні підходи до вивчення поведінки веб-користувачів базуються на загальних підходах до машинного навчання та інтелектуального аналізу даних [2], протягом останнього десятиліття розвивається досить інша точка зору. Нові моделі, засновані на нейрофізіологічній теорії прийняття рішень були застосовані до процесу вибору зв'язку. Ці моделі мають два етапи: навчання і моделювання. У першому параметри моделі пристосовуються до даних користувача. У другому сконфігуровані агенти моделюються в межах веб-структури для відновлення очікуваної поведінки. Основна відмінність від підходу до машинного навчання полягає в тому, що модель не залежить від структури та змісту веб-сайту. Крім того, агенти можуть зіткнутися з будь-якою сторінкою і вирішити, яку посилання прослідкувати (або залишити веб-сайт). Ця важлива характеристика робить такі моделі придатними для сильно динамічних веб-сайтів. Інша важлива відмінність полягає в тому, що ці моделі мають сильну теоретичну основу, побудовану на фізичному явищі, чиї модельні рівняння походять від

спостереження явищ. Підходи традиційних фізичних моделей є більш загальними, але для аналізу поведінки веб-користувачів пропозиція ґрунтується на конкретній сучасній теорії прийняття рішень про мозок.

Нові моделі поведінки користувачів Інтернету пов'язані з рівнями нейронної активності мозку (NAL) певних областей мозку з дискретним набором можливих варіантів. Наприклад, LCA (Leaky Computing Accumulator) використовується для аналізу еволюції NAL (Xi) відповідно до стохастичного рівняння під час процесу прийняття рішення агентом, поки одна зі значень NAL не досягне заданого порогу. Вона описує нейронну активність різних областей мозку під час розв'язання предмета рішення за допомогою стохастичного процесу, який розвивається до тих пір, поки активність не досягне заданого порогу, який викличе рішення. Такий клас стохастичних процесів, що застосовуються до прийняття рішень, експериментально вивчався протягом майже сорока років. У цьому контексті веб-користувач стикається з рішенням, яке посилання буде дотримуватися відповідно до його власних цілей, і цей процес повторюється знову для кожної відвідуваної сторінки до виходу з веб-сайту. Потім веб-користувач стикається з набором дискретних рішень, які відповідають вибору гіперпосилання.

Модель LCA імітує сеанс штучного веб-користувача, оцінюючи послідовності сторінок користувача і, крім того, визначаючи час, необхідний для вибору дії, наприклад, залишивши сайт або перейшовши на іншу веб-сторінку. Експерименти, проведені з використанням штучних агентів, які поводяться таким чином, підкреслюють подібність між результатами штучних агентів і реальною моделлю поведінки веб-користувачів. Крім того повідомляється, що продуктивність штучних агентів має подібну статистичну поведінку для людини. Якщо веб-сайт не змінюється, то набір відвідувачів залишається незмінним. Цей принцип дозволяє прогнозувати зміни в шаблоні доступу на веб-сторінки, пов'язані з невеликими змінами на веб-сайті, які зберігають семантику. Поведінка веб-користувача може бути передбачена шляхом моделювання, а потім сервіси можуть бути оптимізовані. Інші дослідження моделей [3] безпосередньо стосуються методів кластеризації загального призначення.

Такий аналіз поведінки веб-користувачів вимагає нетривіальної стадії попередньої обробки даних, щоб отримати послідовність веб-сторінок (сеансів) для окремих відвідувачів, текстовий контент і структуру гіперпосилання веб-сайту. Розроблені нові алгоритми на основі цілочисельного програмування, які використовуються для оптимального вилучення веб-сесій користувачів. Традиційно наступний крок полягає у застосуванні методів інтелектуального аналізу даних для виявлення та вилучення шаблонів поведінки веб-перегляду користувачів. Крім того, необхідно забезпечити якість даних, оскільки калібрування веб-моделей користувачів є чутливим до набору даних.

Використана література:

1. Velasquez, J.D., Palade, V.: A knowledge base for the maintenance of knowledge extracted from web data. Knowledge Based Systems Journal 20(3), 238–248 (2007);

2. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (2002);

3. Abraham, A., Ramos, V.: Web usage mining using artificial ant colony clustering and genetic programming. In: Procs. of the 2003 IEEE Congress on Evolutionary Computation (CEC 2003), pp. 1384–1391 (2003);

4. Kosala, R., Blockeel, H.: Web mining research: A survey. SIGKDD Explorations: Newsletters of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining 1(2), 1–15 (2000);

5. Anderson, C.: Wired Magazine, Editorial (June 2008).

Іванов О.В. Прогнозування поведінки користувачів веб-орієнтованих систем на основі аналізу серверних логів. Міжнародна наукова інтернет-конференція "Інформаційне суспільство: технологічні, економічні та технічні аспекти становлення (випуск 37)" / Збірник тез доповідей: випуск 37 (м. Тернопіль, 2 квітня 2019 р.). – Тернопіль. – 2019. ISSN 2522-932X. С. 23-25.

Передбачення намірів користувача щодо певного продукту або категорії, заснованої на взаємодії з веб-сайтом, має вирішальне значення для сайтів електронної комерції та мереж показу оголошень, особливо для таргетингу. Відстежуючи пошукові моделі користувачів, онлайн-торговці можуть краще розуміти їх поведінки та наміри.

Перш ніж приступити безпосередньо до аналізу, необхідно розібратися з типами доступних даних. Для цього розглянемо файли журналу веб-сервера - веб-логи.

Для кожного запиту браузера до веб-сервера відгук генерується автоматично, і всі відомості заносяться в веб-лог – текстовий файл з роздільниками в кодуванні ASCII. У серверних логах дані можуть існувати з домішками різних шумів. Кроки попередньої (див. рис. В.1) обробки необхідні для видалення шумів, порожніх і зайвих даних.

Під час попередньої обробки клієнт отримує веб-посилання, що найчастіше використовуються користувачами, сторінки HTTP-запиту, а також все супутню інформацію яка зберігається з запитами відвідувачів.

Використовуючи найчастіші веб-посилання, ми передбачаємо поведінку користувачів та визначаємо що саме переглядає наш відвідувач перебуваючи на сайті.

Існує велика кількість методів для прогнозування поведінки веб-користувачів. Існують чотири основні технології видобутку, які можна застосувати до журналів веб-сервера:

− на основі пошуку послідовних шаблонів (Sequential-pattern) [1]. Дозволяє виявляти тимчасово впорядковані шаблони доступу;

− на основі правил асоціації знаходить співвідношення між типом веб-сторінок;

− кластерне групування. Групи користувачів з подібними характеристиками;

− на основі класифікації: Групи користувачів у попередньо визначені класи на основі їх характеристик.

Рисунок В.1 – Структура системи

Аналіз за допомогою правил асоціації є імперативним методом дослідження. Є багато алгоритмів застосовується для аналізу поведінки користувачів. Найбільш поширеними є Apriori, Apriori TID, STEM, DIC, алгоритм розділів, Eclat і FPGrowth.

Алгоритм Apriori [3] – пошук асоціативних правил, які генеруються на основі всіх частих предметних наборів, виявлених в транзакционной базі даних, і задовольняють заданим рівнем підтримки та достовірності. Даний метод дозволяє скоротити простори пошуку завдяки властивості апріорність. Воно стверджує, що якщо предметний набір Z не є частим, то додавання до нього деякого нового предмета A робить його таким. Іншими словами, якщо Z не є частим, то і Z + A також не їм. Але вузьким місцем в алгоритмі apriori є процес генерації кандидатів в популярні предметні набори. Таким чином, обчислювальні і тимчасові витрати, які потрібні на їх обробку, можуть бути неприйнятними. Крім цього, алгоритм a priori вимагає багаторазового сканування бази даних транзакцій, а саме стільки разів, скільки предметів містить найдовший предметний набір.

Одним з найбільш ефективних процедур пошуку асоціативних правил є алгоритм, який отримав назву Frequent Pattern-Growth (алгоритм FPG) [2], що можна перевести як «вирощування популярних (часто зустрічаються) предметних наборів». Він дозволяє не тільки уникнути витратною процедури генерації кандидатів, але зменшити необхідне число проходів БД до двох. В основі методу лежить попередня обробка бази транзакцій, в процесі якої ця база даних перетворюється в компактну деревоподібну структуру, яка називається Frequent-Pattern Tree – дерево популярних предметних наборів (звідки і назва алгоритму). Надалі для стислості будемо називати цю структуру FP-дерево.

На етапі аналізу моделі (pattern analysis stage) виконується інтерпретація отриманих результатів.

Література:

1. Agrawal, R. and Srikant, R. 1995. Mining sequential patterns, P. S. Yu and A. S. P. Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, 3, 14;

2. Srivastava, J. et al. (2000). Web usage mining: Discovery and applications of usage patterns from Web data, ACM SIGKDD Explorations, 1(2);

3. R. Mishra, A. Choubey, “Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol 2, 2012.

Іванов О.В. Попередня обробка даних для аналізу поведінки користувачів веб-орієнтованих систем. Прикладні наукові розробки та теоретичні дослідження XXI століття: зб. наук. праць «ΛΌГОΣ» з матеріалами міжнар. наук.-практ. конф., м. Вінниця, 15 квітня, 2019 р. Вінниця: ГО «Європейська наукова платформа», 2019. ISBN 978-617-7171-80-4. С 72-74

Дані для аналізу поведінки веб-користувачів отримані з джерел таких як наприклад журнал веб-сервера, потребують попередньої обробки.

На етапі очищення даних спочатку видаляються глобальні та локальні шуми. Глобальні шуми включають дзеркальні сайти, дубльовані веб-сторінки, попередні

версії веб-сторінок та шумні слова. Місцеві шуми [1] включають нерелевантні пункти на веб-сторінці, такі як банерна реклама, навігаційна довідка, прикраси, графічні та відео формати.

Код статусу HTTP розглядається в наступному процесі очищення. Перевіряючи поле стану кожного запису в журналі, записи з кодом стану понад 299 або менше 200 видаляються, оскільки дають не успішну відповідь серверу.

Видалення записів, що не відображають активність користувача. Веб-боти в автоматичному режимі переглядають безліч різних сторінок в мережі [3]. Їх поведінка сильно відрізняється від людського, і вони не представляють інтересу з точки зору аналізу використання веб-ресурсів. Можна застосувати фільтрацію рядків користувача-агента, часто в їх ім'я може бути включений URL або e-mail адресу. Це мабуть найпростіший, але найменш надійний спосіб виявити, чи є він користувачем чи ні.

Багато ботів як правило, підміняють агенти користувачів, а деякі роблять це з поважних причин (тобто вони хочуть лише сканувати мобільний контент), а інші просто не хочуть бути ідентифіковані як боти. Ще гірше, деякі боти підміняють легітимні ботові агенти, такі як агенти користувача google, microsoft, lycos та інших сканерів, які зазвичай вважаються ввічливими. Далі, якщо швидкість перегляду перевищує поріг, ці запити також видаляються. За допомогою реалізації вищезазначених методів очищаються оригінальні файли журналів. Близько 50-60% невідповідних записів видаляються, що сприяє більш якісному результату аналізу поведінки веб-користувачів в подальшому.

Визначення кожного окремого користувача. Більшість порталів в мережі Інтернет доступні анонімним користувачам. Можна застосовувати інформацію про зареєстрованих користувачів, доступні файли cookie для визначення кожного користувача.

Ідентифікація сеансу користувача – це процес сегментації журналу активності кожного користувача на сеанси, кожен з яких являє собою один візит на сайт. Мета евристики сеансу полягає в тому, щоб відновити з даних потоку кліків

фактичну послідовність дій виконаних одним користувачем під час одного відвідування сайту.

Ідентифікація користувача за IP-адресою. Використовується для присвоєння унікальної адреси пристроям (комп'ютеру, принтерам і т.д.), що беруть участь у мережі. IP-адреса записується в журнал, коли користувач потрапляє на сторінку. Цю адресу можна використовувати для розрізнення різних користувачів. Але у випадку проксі-сервера, коли багато користувачів запитують певну сторінку, сервер веб-сайту реєструє той самий IP-адресу (IP-проксі-сервер) у лог-файл. Практично різні користувачі отримують доступ до цієї сторінки. Кешування також створює проблему для ідентифікації унікального користувача. Всякий раз, коли користувач намагається отримати доступ до попередньо переглянутої сторінки, сторінки браузера відображаються з локального кешу, і в журнал не входить запис.

Сеанс може бути ідентифікований атрибутом реферера в розширеному форматі журналу. Припустимо, що X і Y є двома запитами на послідовні сторінки одним і тим самим користувачем і (сеансом), якщо посилання на Y було викликано раніше в цій сесії S, тоді Y буде додано в сеанс S, в іншому випадку нова сесія створюється з Y як перша запитувана сторінка.

Ідентифікація користувача топологією сайту. Цей метод використовує структурну топологію веб-сайту [2] для ідентифікації унікального користувача. Припустимо, що користувач запитує сторінку яка не доступна через попередньо запитані сторінки, тоді він розглядається як новий користувач. Це можна зробити, використовуючи атрибут referrer розширеного формату журналу та інформацію про посилання з топології сайту. Деякі ситуації коли цей підхід призводить до плутанини це якщо користувач робить запит використовуючи сторінки з закладками які не підключені через посилання.

Розглянуті методи предобработки специфічні виключно для даних веб-логів. Однак це не означає, що відомості вже готові до використання і побудови моделей. Далі необхідно провести звичайні кроки обробки, а саме: оцінка якості даних, відновлення пропущених значень, виявлення аномальних значень, нормалізація.

Список використаних джерел:

1. Nithya.P and Dr.P.Sumathi., 2012, “Novel PreProcessing Technique for Web Log Mining by Removing Global Noise and Web Robots”, 2012 National Conference on Computing and Communication Systems 978-1-4673-1953-9/12 © 2012 IEEE;

2. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan (2000), Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Vol.1 Page(s): 12-23;

3. P.-N. Tan, V. Kumar (2000) Modeling of web robot navigational patterns, in: WEBKDD Web Mining for Ecommerce Challenges and Opportunities, Second International Workshop.