Кластеризація Текстових Даних для Підвищення Ефективності Інформаційного Пошуку в Мережі

Денис Блох
Кафедра Програмної інженерії
Харківський Національний університет радіоелектроніки
Харків, Україна
denys.blokh.cpe@nure.ua

Володимир Кобзєв
Кафедра Програмної інженерії
Харківський Національний університет радіоелектроніки
Харків, Україна
volodymyr.kobziev@nure.ua

Textual Data Clustering to Increase the Effectiveness of Information Search on the Web

Denys Blokh
dept. of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
denys.blokh.cpe@nure.ua

Volodymyr Kobziev
dept. of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
volodymyr.kobziev@nure.ua

***Анотація*—Розглядається застосування методів Data Mining для аналізу текстової інформації з баз даних інтернет-ресурсів з метою підвищення ефективності маркетингових кампаній. Особлива увага приділяється кластеризації текстових даних і створенню семантичного ядра, що дозволяє покращити стратегії залучення аудиторії в онлайн-середовищі та оптимізувати маркетингові зусилля.**

***Abstract*—The application of Data Mining methods for the analysis of textual information from databases of Internet resources in order to increase the effectiveness of marketing campaigns is considered. Special attention is paid to the clustering of textual data and the creation of a semantic core, which allows to improve strategies for engaging the audience in the online environment and optimize marketing efforts.**

Ключові слова—інтернет-маркетинг; оптимізація контенту; кластеризація; Data Mining; алгоритм k-середніх

Keywords—internet marketing;content optimization; clustering; Data Mining; k-means algorithm

I. Вступ

Центральною темою даної роботи є використання кластеризації текстової інформації та створення семантичного ядра в контексті інтернет-маркетингу. Дане дослідження спрямоване на розробку стратегій для залучення аудиторії в онлайн-середовищі та підвищення ефективності маркетингових зусиль. Основна мета будь-якого інтернет-ресурсу полягає у створенні релевантного контенту, будь то текстові, графічні, аудіо або відео дані, і максимальному поширенні цього контенту серед потенційної аудиторії з метою монетизації. Пошук в інтернеті, як правило, здійснюється на основі текстових даних, які описують цей контент [1].

В умовах сучасної високої конкуренції в онлайн-середовищі інтернет-ресурси постійно шукають способи підвищення ефективності своїх маркетингових кампаній. Одним із перспективних підходів є використання методів Data Mining для аналізу текстової інформації, яка міститься в базі даних інтернет-ресурсу. Ця робота присвячена вивченню застосування кластеризації текстових даних та створення семантичного ядра для оптимізації маркетингових зусиль і залучення користувачів.

Інтернет-маркетинг сьогодні стає все більш конкурентним, що вимагає від компаній постійного вдосконалення своїх підходів до створення та розповсюдження контенту. Основна мета будь-якого інтернет-ресурсу полягає у створенні релевантного контенту, який привертає увагу цільової аудиторії та сприяє монетизації. Незалежно від форми контенту (текстової, графічної, аудіо чи відео) пошук інформації здебільшого здійснюється на основі текстових описів.

В умовах сучасної конкуренції інтернет-ресурси мають постійно шукати нові способи підвищення ефективності своїх маркетингових кампаній. Одним з найбільш перспективних підходів є використання методів Data Mining для аналізу текстових даних, які містяться у базах даних інтернет-ресурсів. Дана робота зосереджена на вивченні застосування кластеризації текстових даних та створенні семантичного ядра з метою оптимізації маркетингових зусиль та залучення користувачів.

Кластеризація текстової інформації дозволяє групувати схожі документи та створювати семантичні ядра, що допомагають покращити релевантність контенту. Це, у свою чергу, сприяє підвищенню видимості ресурсу у пошукових системах та залученню більшої кількості відвідувачів.

Дана робота має на меті показати, як кластеризація текстових даних може бути ефективно використана для підвищення ефективності інтернет-маркетингу, шляхом

створення релевантного та цільового контенту, що відповідає потребам користувачів.

II. Основний матеріал

Метою роботи є розробка і впровадження методів Data Mining для аналізу текстової інформації з метою оптимізації маркетингових кампаній. Основними завданнями є:

- вивчення сучасних технологій Data Mining та їх застосування в аналізі текстових даних,
- розробка методики кластеризації текстових даних для виділення релевантних тематичних груп,
- створення семантичного ядра на основі кластеризованих даних,
- оцінка ефективності запропонованих підходів у контексті інтернет-маркетингу.

Data Mining — це процес виявлення патернів у великих наборах даних з використанням методів машинного навчання, статистики та штучного інтелекту. Основою Data Mining [2] є концепція шаблонів, які відображають закономірності в підвибірках даних, дозволяючи знаходити автономні взаємозв'язки без попередніх припущень про дані. Це робить Data Mining потужним інструментом для аналізу великих обсягів інформації та виявлення неочевидних патернів, які можуть бути пропущені при використанні традиційних статистичних методів.

A. Кластеризація

Кластеризація є однією з основних технік Data Mining, що використовується для групування схожих об'єктів даних у кластери. Цей підхід дозволяє аналітикам виділяти групи схожих об'єктів, аналізувати їхні особливості та будувати окремі моделі для кожної групи. Кластеризація особливо важлива як один з етапів аналізу даних і побудови завершених аналітичних рішень.

Кластеризація забезпечує ефективне групування аналогічних об'єктів даних, завдяки чому об'єкти в межах одного кластера є подібними, але відрізняються від об'єктів інших кластерів. Це сприяє більш глибокому розумінню структури даних і підвищенню точності аналітичних моделей.

Одним із найефективніших методів кластеризації є алгоритм K-середніх (k-means) [3]. Він формує кластери з високим ступенем внутрішньокласової схожості, зберігаючи при цьому низьку схожість між кластерами.

Основні кроки алгоритму K-середніх:

- вибір фіксованої кількості кластерів (K),
- ініціалізація центроїді. Початкові координати центроїдів (центрів тяжіння кластерів) вибираються випадковим чином або з використанням певних спеціальних методів,
- призначення об'єктів до кластерів. Кожний об'єкт даних призначається до кластеру з найближчим центроїдом,
- оцінка схожості. Схожість між об'єктами в межах одного кластера оцінюється середньою відстанню його об'єктів до центроїду,
- оптимізація кластерів. Алгоритм повторює процес призначення об'єктів до кластерів і перерахування центроїдів до тих пір, поки не досягне стабільності кластерів (стабільного розподілу об'єктів).

Кластеризація є важливою технікою в Data Mining, яка дозволяє ефективно групувати схожі об'єкти даних для подальшого аналізу та моделювання. Алгоритм K-середніх є одним з найефективніших методів кластеризації, що забезпечує високу точність і швидкість обчислень. Використання кластеризації сприяє покращенню розуміння структури даних та підвищенню якості аналітичних рішень.

B. Оптимізація контенту

Основною метою будь-якого інтернет-ресурсу є створення релевантного контенту незалежно від його форми — текстової, графічної, аудіо або відео. Цей контент повинен бути максимально поширеним серед потенційної аудиторії для досягнення високого рівня монетизації. Оптимізація контенту базується на аналізі текстових даних, що дозволяє отримувати глибокі уявлення про найбільш обговорювані теми серед аудиторії та краще зрозуміти інтереси та потреби цільової аудиторії.

C. Практичне застосування

Кластеризація за словоформою базується на об'єднанні запитів на основі їх структурної або граматичної схожості. Незважаючи на різний контекст, різні запити можуть мати однакові словоформи в основі. Для зручності користувачів доцільно розділити запити в кластери за напрямками: інформаційні, навігаційні та транзакційні, що дозволить швидко знаходити релевантну інформацію.

На існуючому інформаційному ресурсі проведено експериментальні заходи з аналізу запитів. Обсяг даних про користувачів ресурсу становить понад 500,000 активних користувачів, тому застосовано підхід Big Data для обробки даних. Через великий обсяг інформації мануальна обробка є недоцільною, тому використовувався сервіс автоматизованої кластеризації Serpstat [4].

Процес кластеризації. Запити об'єднуються в один кластер, якщо вони мають високі шанси потрапити на першу сторінку пошукової системи Google. Якщо два запити можна інтегрувати в одну статтю, їх не розносять по різних кластерах. Для кластеризації, враховуючи відому кінцеву кількість кластерів (кількість рубрик проекту), використано алгоритм k-середніх.

Аналіз текстових даних дозволив створити ядро з 4500 ключових запитів, серед яких відібрано лише ті, що увійшли в топ пошукових запитів. Як результат залишилось 350 запитів, поділених на 10 кластерів. Під ці запити адаптована текстова інформація інтернет-ресурсу.

Дані підготовлені шляхом відбору та підготовки відповідних ознак. Аналіз кластерів дозволив виявити важливі відмінності та недоліки в організації контенту (створюваний контент не завжди відповідав потребам користувачів). Це спонукало до реорганізації контенту та маркетингової стратегії, щоб більш точно відповідати очікуванням аудиторії та підвищити її зацікавленість.

Роботи з оптимізації інтернет-ресурсу проводилися протягом 2 місяців, суттєвий результат був відчутний через 1 місяць після внесених змін. У підсумку були створені шаблони, які раніше не розглядалися взагалі. Результат досягнуто завдяки створенню нових форм і шаблонів подачі інформації, які краще відповідають потребам користувачів, оптимізації маркетингових витрат. Графік ефективності ресурсу за весь рік (Рис. 1) ілюструє стійке зростання зацікавленості після проведення оптимізації контенту та маркетингової стратегії, що свідчить про успішність внесених змін. Порівняння ефективності проведених робіт наведено у таблиці 1.

Рис. 1.Статистика переглядів за рік

ТАБЛИЦЯ 1 ПОРІВНЯННЯ РЕЗУЛЬТАТІВ ОПТИМІЗАЦІЇ РЕСУРСУ ЗА РАХУНОК КЛАСТЕРИЗАЦІЇ ІНФОРМАЦІЇ

Кількісний показник	***До***	***Після***
Переходів на ресурс за добу (людин)	22000	90000
Загальна тривалість перебування всіх користучів на ресурсі за добу (годин)	3200	15600
Кількість людин, які підписувалися на оновлення за добу (люин)	132	810

У сучасних умовах об'єм інформації стає настільки великим, що авторам важко зрозуміти, який контент найбільш цікавий та корисний для їхньої аудиторії. З цієї причини кластеризація текстових даних може стати потужним інструментом для оптимізації вибору контенту та підвищення його релевантності для аудиторії.

Важливо зазначити, що головним джерелом трафіку залишився пошук в Google, альтернативним джерелом трафіку виступала контекстна реклама. Після індексації пошуковою системою внесених на ресурсі змін, відбувся стрімкий стрибок відвідування ресурсу при тому, що статистика переходів з контекстної реклами не змінилася (Рис. 2).

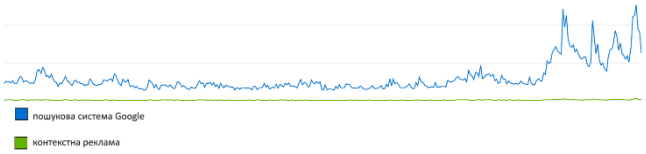

Рис. 2. Статистика переходу на ресурс через пошукову систему та контекстну рекламу

При цьому відбулися суттєві зміни вікового показника користувача, з`явився новий сегмент користувачів віком від 30 до 60 років, який раніше був майже відсутній серед користувачів ресурсу (Рис. 3).

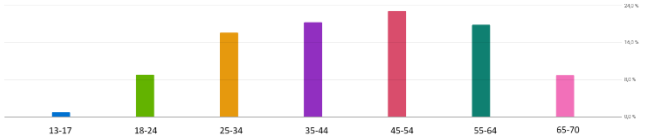

Рис. 3. Діаграма віку відвідувача ресурсу після оптимізації

Подальший аналіз і застосування на практиці кластеризованих даних окремо за інформацією ресурсу і про її користувачів показали високу ефективність при оптимізації маркетингової стратегії та організації створення контенту. Кластеризація розкрила важливі відмінності та недоліки в організації контенту.

III. ВИСНОВКИ

Аналіз інформації виявив важливі відмінності та недоліки в організації контенту, що спонукало до його переорганізації та вдосконалення маркетингової стратегії. Впровадження змін дозволило досягти суттєвого зростання зацікавленості користувачів, що в свою чергу призвело до значного підвищення монетизації на 120%. Ці результати підкреслюють ефективність використання кластеризації та Data Mining у контексті оптимізації контенту та маркетингових зусиль.

Кластеризація текстових даних є ефективним інструментом для оптимізації контенту та підвищення його релевантності для аудиторії. Її застосування в інтернет-маркетингу сприяє покращенню стратегії приваблення користувачів, що підтверджується успішними результатами експериментів та суттєвим зростанням активності користувачів.

ЛІТЕРАТУРА REFERENCES

[1] Serpstat. (2021). Serpstat: SEO Platform for Professionals. Retrieved from https://serpstat.com

[2] Han, Jiawei. Data mining: concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed., Morgan Kaufmann Publishers is an imprint of Elsevier, 2012, 740p.

[3] Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., & McLachlan, G. J. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37.

[4] Microsoft. (2018). Microsoft Clustering Algorithm. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-clustering-algorithm?view=sql-server-2017