Контент-аналіз текстових масивів даних

Євген Кондратєв[1], Вікторія Висоцька[2]

Information Systems and Networks Department, Lviv Polytechnic National University, 12 S. Bandera street, Lviv, UKRAINE, E-mail: [1]kondratyev.yevhen@gmail.com, [2]victana@bk.ru

Abstract – From the perspective of a systemic approach, the principles of applying information resources processing for content lifecycle implementation made the development of methods for the commercial content formation possible. An integrated method of commercial content formation for the time and resources reduction of content production is developed. This makes it possible to create a means of information resources processing and implement subsystem of automatically generated content.

Keywords – text, content, commercial content, content analisis, content monitoring.

Вступ

Негативні чинники у формуванні контенту ускладнюють процес пошуку необхідних даних при скануванні різних джерел інформації. Збільшення фізичного обсягу та змінна актуальності/динаміки потоків контенту (систематичне та не регулярне оновлення) призводить до виникнення дублювання, інформаційного шуму та надмірності результатів пошуку контенту. Охоплення та узагальнення великих динамічних потоків контенту, які неперервно генерують в Інтернет-джерелах, вимагає якісно нових методів/підходів пошуку як контент-моніторинг. Вхідною інформацією для нього є текст на природній мові як послідовність символів, вихідна інформація – це таблиці розділів, речень і лексем аналізованого тексту. Контент-моніторинг є засобом автоматизації знаходження важливих складових в потоках контенту. Це змістовний аналіз потоків контенту для отримання необхідних якісних/ кількісних зрізів на протязі наперед не визначеного проміжку часу.

Основна частина

Складовою контент-моніторингу є контентний пошук та контент-аналіз тексту. Контент-аналіз призначений для пошуку контенту в масиві даних за змістовими лінгвістичними одиницями. Одиниця рахунку є кількісною мірою одиниці аналізу, що дозволяє реєструвати частоту (регулярність) появи ознаки категорії аналізу в тексті (кількість певних слів або їх поєднань, рядків, друкованих знаків, сторінок, абзаців, авторських аркушів, площа тексту тощо).

Алгоритм 1. Контент-аналіз текстового комерційного контенту.

Етап. 1. Визначення набору критеріїв ддля текстового комерційного контенту.

Крок 1. Формування набору критеріїв як тип джерела (форум, електронна пошта, Інтернет-газета, чат, Інтернет-журнал); тип контенту (стаття, електронний лист, баннер, коментарій); учасники комунікації (відправник, одержувач, реципієнт).

Крок 2. Визначення розміру (мінімальний обсяг або довжина), частоти появи, способу/місця розповсюдження та час появи контенту.

Крок 3. Фільтрування згідно сформованого набору критеріїв контентоного потоку та зберігання ідентифікованого релеватного контенту.

Етап. 2. Контент-аналітичний відбір. Формування вибіркової сукупності контенту за критеріями обмеженої вибірки з більшого масиву.

Етап. 3. Виявлення змістовних одиниць аналізу текстового контенту (словосполучення, речення, тема, ідея, автор, персонаж, соціальна ситуація, частина тексту, кластеризованна за змістом категорії аналізу). Вимоги до вибору лінгвістичної одиниці аналізу: достатньо велика для інтерпретації значення; достатньо мала, щоб не інтерпретувати багато значень; легко ідентифікується; кількість одиниць достатньо велика для проведення вибірки.

Етап. 4. Виділення одиниць рахунку аналізу текстового контенту.

Крок 1. Якщо одиниці рахунку збігаються з одиницями аналізу, то знаходять частоти появи виділеної змістовної одиниці, інакше перейти до кроку 2.

Крок 2. Модератор на основі аналізованого контену висуває одиниці рахунку, наприклад, протяжність текстів; площа тексту, заповнена змістовними одиницями; кількість рядків (абзаців, знаків, колонок тексту); розмір/вид файлу; кількість рисунків з певним змістом/сюжетом тощо.

Етап. 5. Порівняння змістовних одиниць аналізу з одиницями рахунку.

Крок 1. Класифікація за угрупованнями із оціненням ваги змістовних категорій в загальному обсязі тексту. Класифікатором є загальна таблиця, в яку зведені всі категорії аналізу і одиниці аналізу. Фіксують одиниці виразу категорій.

Крок 2. Статистичні розрахунки зрозумілості та атрактивності контенту.

Етап. 6. Розроблення інструменту контент-аналізу.

Крок 1. Створення закодованого протоколу контенту для компактності подання даних та швидкого порівняння результатів аналізу різного контенту.

Крок 2. Заповнення протоколу контенту властивостями (автор, час, обсяг тощо).

Крок 3. Заповнення протоколу контенту підсумками його аналізу (кількість вживання в ньому певних одиниць аналізу і висновки щодо категорій аналізу). Протокол кожного контенту заповнюється на основі підрахунку даних всіх його реєстраційних карток.

Етап. 7. Розроблення таблиці контент-аналізу. Тип таблиці визначають у вигляді системи скоординованих і субординованих категорій аналізу: кожна категорія (питання) передбачає ряд ознак (відповідей), за якими квантифікується зміст тексту.

Етап. 8. Розроблення кодувальної матриці контент-аналізу.

Крок 1. Якщо обсяг вибірки ≥ 100 одиниць, то аналізується набір матричних листів, інакше виконати крок 2.

Крок 2. Якщо вибірка < 100 одиниць, то проводиться двовимірний аналіз. В цьому випадку для кожного контенту формується кодувальна матриця.

Етап. 9. Проведення аналізу тексту згідно створених кодувальних матриць.

Етап. 10. Інтерпретація результатів. Виявляють і оцінюють характеристики контенту на основі статистичного набору підрахованих коефіцієнтів за період часу на категорію.

Висновок

Застосування контент-аналізу при моніторингу Інтернет-джерел даних дозволяє автоматизувати процес знаходження найбільш важливих складових в потоках контенту при відборі даних з цих джерел. Це усуває дублювання контенту, інформаційний шум, паразитичний контент, надмірність результатів пошуку тощо. Цей метод застосовують в подальших етапах формування контенту для отримання більш точного релеватного результату – створення унікального контенту, який користується попитом.