УДК 004.9

ЗАСТОСУВАННЯ КЛАСТЕРНОГО АНАЛІЗУ ДЛЯ ВІДСІЮВАННЯ ПОШУКОВОГО СПАМУ

Зосімов В.В.

Миколаївський національний університет ім. В.О. Сухомлинського,
вул. Нікольська, 24, Миколаїв, 54030, Україна

zosimovvv@bk.ru

Представлений підхід до відсіювання певного виду пошукового спаму з результатів видачі пошукових систем. Відсіювання здійснюється за рахунок об'єднання неунікальної інформації в кластери з подальшим їх вилученням з видачі пошукових систем. Для групування веб-сторінок запропоновано використовувати метод кластерного аналізу на основі моделей, побудованих із застосуванням індуктивних алгоритмів.

Ключові слова: Інтерфейс, кластерний аналіз, онтології, індуктивне моделювання, веб-сторінка.

This paper presents an approach to sifting out a certain type of search spam from the search engines results. Dispersion is carried out by combining non-unique information in clusters with their subsequent exclusion from the search engines. It was suggested to use the cluster analysis method based on models built using inductive algorithms to group similar web pages.

Keywords: Interface, cluster analysis, ontology, inductive modeling website.

Представлен подход к отсеиванию определенного вида поискового спама из результатов выдачи поисковых систем. Отсеивание осуществляется за счет объединения неуникальной информации в кластеры с последующим их исключением из выдачи поисковых систем. Для группировки веб-страниц предложено использовать метод кластерного анализа на основе моделей, построеных с применением индуктивных алгоритмов.

Ключевые слова: Интерфейс, кластерный анализ, онтологии, индуктивное моделирование, веб-страница.

Вступ

В даний час велика кількість веб-ресурсів створюється з метою залучення якомога більшої кількості відвідувачів і конвертації дій користувача в реальні гроші. Прибуток автори сайтів отримують за рахунок активності залучених на їх веб-ресурс користувачів. Це може бути перехід по контекстній рекламі, перегляд рекламних банерів і відеороликів, перехід на сайти продавців різних товарів через партнерські програми, тощо.

Для залучення нових відвідувачів і утримання вже існуючих на веб-ресурсі необхідно регулярне додавання нового якісного і цікавого контенту. Такий спосіб розвитку веб-ресурсів використовує лише невеликий відсоток авторів. Це обумовлено тим, що на збір даних і створення якісного текстового,

графічного або відео-контенту потрібно багато часу. Крім часу необхідні певні професійні навички.

Більшість авторів блогів не створюють унікальний контент для своїх веб-ресурсів. Вони отримують його шляхом невеликої переробки вже існуючих популярних матеріалів. Витрачаючи при цьому мінімум часу і зусиль. У цьому випадку вміст веб-ресурсу, як правило, несе мінімум корисної інформації. При створенні подібних веб-ресурсів вміст підганяється під критерії, необхідні для успішного SEO-просування, часто це негативно впливає на якість інформації та зручність для користувача. При цьому вміст є «рерайтом» (викладанням у вільній формі вже існуючого матеріалу) деякого унікального тексту, без вказівки посилання на джерело. Ще одним мінусом є те, що рерайт здійснюють не фахівці з конкретної предметної області, а люди, які вміють зв'язно вести викладання матеріалу, не вникаючи в суть питання, що часто призводить до появи в тексті неточностей, які вводять користувача в оману.

Іноді кількість таких «клонів» досягає сотень і тисяч примірників. Найчастіше такий підхід застосовується в новинних сайтах і блогах. Виявлення оригінального тексту, як правило є дуже складним завданням.

Всі результати пошуку, які не містять унікальну інформацію, а тільки непрофесійно створені клони, є пошуковим спамом, тому що не несуть для користувача ніякої корисної інформації.

У деяких випадках на першій сторінці результатів пошуку 10 з 10 джерел будуть рерайтом одного документа. Пошукові системи безперервно вдосконалюють алгоритми ранжирування, але виявлення в результатах пошуку неунікальні інформації все ще залишається однією з найактуальніших завдань.

1. Можливості застосування кластерного аналізу при пошуку «рерайту» вихідного тексту

Кластерний аналіз успішно застосовується для вирішення завдань підвищення ефективності пошуку інформації в інтернеті.

Формальна постановка задачі кластеризації.

Нехай X — множина об’єктів, Y — множина номерів (імен, міток) кластерів. Задана функція відстані між об’єктами *p(x,x')*. Існує скінченна навчальна вибірка об’єктів $X^m = \{x_1, \ldots, x_m\} \subset X$. Потрібно розбити вибірку на непересічні підмножини, які називаються кластерами, так, щоб кожен кластер складався з об'єктів, близьких за метрикою ρ, а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту $x_i \in X^m$ приписується номер кластера y_i.

Алгоритм кластеризації — це функція $a: X \to Y$, яка кожному об’єкту $x \in X$ ставить у відповідність номер кластера $y \in Y$. Множина Y в деяких випадках відома заздалегідь, однак частіше ставиться завдання визначити оптимальне число кластерів, з точки зору того чи іншого критерію якості кластеризації.

Кластеризація відрізняється від класифікації тим, що мітки вихідних об'єктів y_i спочатку не задані, і навіть може бути невідома сама множина Y.[1]

Розв’язання задачі кластеризації принципово неоднозначно, і на те є кілька причин:

• не існує однозначно найкращого критерію якості кластеризації. Відомий цілий ряд евристичних критеріїв, а також ряд алгоритмів, які не мають чітко вираженого критерію, але здійснюють досить розумну кластеризацію «з побудови». Всі вони можуть давати різні результати. Отже, для визначення якості кластеризації потрібен експерт предметної області, який би міг оцінити осмисленість виділення кластерів.

• число кластерів, як правило, невідомо заздалегідь і встановлюється відповідно до деяких суб'єктивних критеріїв. Це справедливо тільки для методів дискримінації, так як в методах кластеризації виділення кластерів йде за рахунок формалізованого підходу на основі заходів близькості.

Результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається експертом. Але варто відзначити, що є ряд рекомендацій до вибору мір близькості для різних завдань [1].

Існує кілька пошукових систем, в яких застосовується кластерний аналіз для більш комфортного представлення результатів пошуку користувачу.

Механізми, що використовують кластерний аналіз забезпечують кращу презентацію результатів пошуку, тому що організують їх в структуру. Цей метод полягає в призначенні певних категорій або тематик документам і результатів пошуку. Поняття кластер означає множину, сукупність, зв'язку або просто групу. Кластеризація спрямована на, наскільки це можливо, сортування результатів пошуку в одну або кілька таких груп, і таким чином отримує групи з усіх результатів.

Умовою успіху є попереднє визначення груп, а також категорій, тематик або шарів, які, в свою чергу, визначаються ключовими словами і професійними поняттями. Для того, щоб документ міг бути призначений групі, він повинен бути правильно класифікований. Реалізація цього наміру не завжди виявляється досить простою. Щоб її зробити, пошукова система читає, після чого досліджує дані і метадані документа. Також аналізує зміст документа на основі статистичних розрахунків (бере до уваги частоту появи букв, складів і слів, порядок фраз, а також довжини слів і пропозицій), або використовує алгоритми лінгвістичного аналізу. Чим точніші ці дані, тим точніше можна виділити документ в певній групі [2].

Однак кластерний аналіз можна застосовувати не тільки для групування результатів, але і для відсіювання деяких видів пошукового спаму.

Ідея застосування кластерного аналізу як методу боротьби з пошуковим спамом полягає в виявленні та групуванні неунікальний текстів у кластери, які можуть бути замінені на один з результатів або видалені з пошукової видачі повністю.

Для визначення того, чи присутній рерайт в пошуковій видачі Google, було проведено ряд експериментів.

Пошукова видача аналізувалася в ручному режимі. Для кожного експерименту були розглянуті перші 50 результатів пошуку.

Результати одного з цих експериментів представлені нижче.

В експерименті брали участь 5 студентів спеціальності «Комп'ютерні науки». Перед ними стояло завдання переглянути перші 5 сторінок пошукової видачі Google і розбити результати пошуку на групи, а також оцінити, чи корисна була інформація, представлена на сайті для вирішення поставленого в пошуковому запиті завдання.

Пошуковий запит: «вставити відео на сайт»

Серед результатів були виявлені наступні групи джерел інформації:

Група 1.

Інструкція по додаванню відео з сайту youtube.com - 7 шт.

Представлена на них інформація поверхнева і за змістом ідентична.

Група 2.

Інструкції по додаванню відео з різних онлайн сервісів, включаючи сайт youtube.com - 23 шт

Представлена на них інформація поверхнева і за змістом ідентична.

Група 3.

Посилання на онлайн довідник з описом тега <video>, який дозволяє вставляти відео на сайт не залежно від джерела - 2шт

Інформація вичерпна і унікальна, її можна вважати корисною.

Група 4.

Онлайн підручники, які надають вичерпну інформацію по темі з докладним описом всіх можливих варіантів - 8 шт.

Цю інформацію можна вважати корисною.

Група 5.

Джерела, що надають інформацію тільки по одному специфічному варіанту, як правило без докладного опису (відео з Facebook, плагін до певної CMS, вузькоспеціалізований онлайн сервіс) - 10 шт

Інформація унікальна.

Перші дві групи сайтів несуть мінімум корисної інформації для користувача, тому замість всіх результатів з цих груп можна показувати користувачеві тільки один з кожної групи, або виключити ці групи з пошуку цілком, так як інформацію представлена на них більш повно і зрозуміло викладена на сайтах з четвертої групи.

2. Підхід до побудови кластерів із застосуванням індуктивних алгоритмів

У розглянутій в статті задачі немає необхідності заздалегідь визначати оптимальне число кластерів. Це число буде різним для кожного пошукового запиту і залежить від кількості першоджерел, кількості результатів, критеріїв кластеризації.

Для якісного розбиття веб-сторінок на групи необхідно побудувати моделі кластеризації. Як інструмент для побудови моделей кластеризації був обраний узагальнений ітераційний алгоритм методу групового урахування аргументів. Цей алгоритм показав високі результати при вирішенні завдань побудови моделей ранжування для пошукових систем [3].

В цій задачі основною проблемою є виявлення характерних ознак, на основі яких буде будуватися модель кластеризації. Для різних результатів пошуку буде доцільно застосовувати окремі ознаки.

Можливим рішенням може бути створення онтологій з описом різних сфер життєдіяльності, до яких може відноситися введений користувачем запит. Кожна онтологія буде містити певний набір ознак кластеризації та моделі кластеризації [4].

Для підвищення ефективності необхідно додати механізми навчання для поновлення ознак кластеризації.

Метод групового урахування аргументів заснований на принципах теорії навчання і самоорганізації, зокрема, на принципі масової «селекції» або самоорганізующимуся направленому переборі всіх можливих варіантів побудови вирішального правила класифікації. Задача побудови вирішального правила в МГУА представляється як задача індуктивної побудови моделі, що ускладнюється в процесі роботи алгоритму.

В роботі розглядається клас задач моделювання, який містить інформацію про n вимірювань m вхідних змінних (ознак) $X[n \times m]$ та однієї вихідної змінної $y[n \times 1]$. Необхідно знайти модель залежності вхід-вихід.

Шукана за допомогою МГУА модель для розглянутої в даній роботі задачі буде представлена в класі підмножин одночленів полінома Колмогорова-Габора:

$$y(x_1,\ldots,x_m) = \theta_0 + \sum_{i=1}^{m} \theta_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} \theta_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \theta_{ijk} x_i x_j x_k + \ldots , \quad (1)$$

де $\theta = \{\theta_0, \theta_i, \theta_{ij}, \theta_{ijk}, \ldots\}$ – вектор коефіцієнтів.

Для визначення вектора коефіцієнтів математичної моделі за вибіркою даних застосовується метод найменших квадратів (МНК).

Задача побудови моделі з вибором її структури і оцінки параметрів зводиться до формування за вибіркою експериментальних даних деякої

множини Ф моделей-кандидатів $f \in \Phi$ різної структури в класі лінійної за параметрами функції (2):

$$\hat{y}_f = f(X, \hat{\theta}_f) \tag{2}$$

і пошуку оптимальної моделі з цієї множини Ф як рішення задачі дискретної оптимізації за умови мінімуму зовнішнього критерію селекції $CR(\cdot)$:

$$f^* = \arg\min_{f \in \Phi} CR(y, f(X, \hat{\theta}_f)), \tag{3}$$

У ролі критерію селекції будемо використовувати критерій регулярності, який заснований на розбитті вибірки на навчальну (А) і перевірочну (В):

$$AR_{B|A} = \left\| y_B - \hat{y}_{B|A} \right\|^2 = \left\| y_B - X_B \hat{\theta}_A \right\|^2 \tag{4}$$

В роботі буде розглядатися узагальнений ітераційний алгоритм (УІА) - гібридний алгоритм МГУА, який об'єднує властивості комбінаторного та багаторядного алгоритмів, виключаючи при цьому їх недоліки, перелічені вище [5].

3. Висновки

Застосування кластерного аналізу для відсіювання неунікальних веб-сторінок дозволить значно підвищити ефективність пошуку інформації за рахунок виведення на перші сторінки пошукової видачі тільки унікальної інформації.

Література

1. http://wikipedia.org
2. http://search.carrotsearch.com/
3. Zosimov V., Stepashko V., Bulgakova O. Inductive building of search results ranking models to enhance the relevance of the text information retrieval. – Proc. of the 26th Intern. Workshop “Database and Expert Systems Applications, 1-4 Sept., Valencia, Spain / Ed. by Markus Spies at al. – Los Alamitos: IEEE Computer Society, 2015. – 316 p. / – P. 291-295. – ISSN: 1529-4188.
4. Antoniou, G., Franconia, E., & van Harmelen, F. (2005). Introduction to Semantic Web Ontology
5. Stepashko V.S., Bulgakova O.S. Generalized iterative algorithm of the group method of data handling // USiM. – 2013. – № 2. – P: 5–18.