УДК 004.42

Левицька Т.О.[1], Чварков М.Д.[2]

ДОСЛІДЖЕННЯ СУЧАСНИХ ТЕНДЕНЦІЙ ВИКОРИСТАННЯ КЛАСТЕРИЗАЦІЇ ПРИ ВИРІШЕННІ ЗАДАЧІ ПЕРСОНАЛІЗАЦІЇ САЙТУ

Дана робота присвячена вирішенню проблем персоналізації користувачів за їх запитами. Аналіз показав, що найкращим методом кластеризації пошукових профілів, в даному випадку, є метод кластеризації CLOPE. Розроблено масштабований алгоритм, що відрізняється великою масштабністю та легкою складністю реалізації

Під час кластеризації відбувається аналіз близьких пошукових профілів користувачів, які звертаються до системи, після чого йому відображаються раніше не переглянуті сторінки, на основі пошукового профілю користувача. На основі розробленої моделі та алгоритму запропонована система персоналізації, яка може інтегруватися з веб-сайтами для підвищення ефективності доступу до релевантної для користувача інформації. Прикладом персоналізації на основі поточних потреб є популярні в даний час рекомендаційні системи, які працюють на основі пошукових профілів користувачів. Відвідувачі, які будуть вперше опинятися на Web сайті або веб-ресурсі будь-якої організації, будуть судити про неї за якістю і значущості опублікованих матеріалів.

Для великих веб-ресурсів існує актуальне завдання швидкої навігаційної і пошукової підтримки користувачів. Його можна вирішити шляхом персоналізації контенту, з урахуванням потреб і особливостями поведінки кінцевого користувача. При персоналізації веб-сторінок буде здійснюватися динамічна зміна вмісту веб-ресурсу під конкретні потреби користувача. У результаті не тільки користувач буде «спілкуватися» з веб-сторінкою, але і сам сайт буде звертатися до будь-якого, який потрапив на сторінку, не як до частини загальної маси, а як до конкретної людини, що має свої особисті інтереси - персонально. Для вирішення питання кластеризації був обраний алгоритм CLOPE, яких призначений для кластеризації великих обсягів даних. В алгоритмі CLOPE, під час роботи зберігається невелика кількість даних по кожному кластеру, при мінімальній кількості сканувань. Метою даної статті є публікація результатів дослідження сучасних тенденцій використання кластеризації при вирішенні задачі персоналізації.

Ключові слова: *персоналізація сайту, веб-ресурс, алгоритм, кластерізація, користувач*

Вступ. Сьогодні організації та велики ресурси переносять свою інформацію і дані на Web-сайти в вигляді рекламної інформаційної продукції або виміряних даних. На Web сайти міститься величезна кількість інформації, яка забезпечує інформаційні потреби багатьох потенційних споживачів. Web технології змінили взаємини між споживачем інформаційної продукції та її творцем - тепер споживачі у багатьох випадках самі шукають дані, які могли б допомогти прийняти правильне рішення, надати свіжу і значиму інформацію.

Тому незалежно від того, використовується Web сайт для реклами своєї діяльності або за допомогою сайту організовується обслуговування даними, зручний і надійний Web сайт

[1] *кан. техн. наук, ДВНЗ «Приазовський державний технічний університет», г. Маріуполь, tlevitiisys@gmail.com*

[2] *бакалавр, ДВНЗ «Приазовський державний технічний університет», г. Маріуполь, suomuro@gmail.com*

абсолютно необхідний. Досвідчений користувач навряд чи буде відвідувати повільно відкривається або погано працює Web сайт. Необхідна доступність і швидка реакція - чи йде мова про локальну корпоративної мережі або про центр даних з величезним трафіком. Відвідувачі, вперше опинилися на Web сайті організації, будуть судити про неї за якістю і значущості опублікованих матеріалів.

Сьогодні майже усі методи та засоби персоналізації орієнтуються на поточні потреби користувачів, що не дає максимальну точність рекомендацій. Саме тому є актуальна задача розробки комбінованої моделі персоналізації і алгоритмів управління контентом сайтів, які враховують постійні і поточні потреби користувачів.

Аналіз останніх досліджень і публікацій. Сьогодні динамічна персоналізація контенту сайту – це обов'язкове правило для успішності та затребуваності веб-ресурсу. Не виключенням є і пошукові системи Google, Yahoo!, Bing, Яндекс, які вже тривалий час надають результати пошукової видачі користувачам в залежності від місця розташування користувача, його мови і, навіть, уподобань користувача [1]. Все це вони здійснюють на підставі історії його попередніх запитів. В своїх сервісах, наприклад, YouTube намагається не просто показати користувачу останні додані відео або аудіофайли, а намагається вгадати, що саме його зацікавить.

У своїх дослідженнях провідні фахівці в цій галузі: Якоб Нільсен та Кара Перніче [2], Хоа Лоранжер [3], Джесс Гарретт [4] і Марі Тахір [5] , неодноразово порушували питання зручності сприйняття інформації на веб-сторінках різними користувачами в залежності від їх статі, віку, а також типу сприйняття інформації (візуали, аудіали, кінестетики). Відносно персоналізації інформації на веб-сайтах для кожного конкретного користувача компанією Яндекс в 2013 році був запущений проект Яндекс.Атом [6], технологія, яка ґрунтується на різних технологіях Яндекса і спрямована на персоналізацію інформації на веб-сайтах для кожного конкретного користувача на підставі інформації , зібраної про нього раніше з різних джерел (сайтів, де код статистики Яндекс.Метрика, панелей браузерів Яндекс і різних сервісів компанії). На даний момент проект зупинений.

Але є певні проблеми з обробкою великих даних отриманих по користувачу для формування рекомендації або динамічного контенту.

Однією з найбільш актуальних проблем обробки великих даних є кластеризація веб-користувачів на основі їхніх спільних властивостей. У статті [7] представлений спосіб визначення подібності інтересів Інтернет-користувачів. Веб-журнали доступу користувачів забезпечують точну і об'єктивну інформацію про відвідувачів. Записи журналу містять IP-адреса веб-користувача, дату і час запиту, URL-адресу запитуваної сторінки, протокол запиту, код повернення сервера із зазначенням статусу обробки запиту і при успішному запиті розмір сторінки. З журналу веб-сервера витягується призначений для користувача шаблон, що складається зі сторінок, які користувач відвідав і витраченого на це часу. Проведені експерименти показали, що запропонований метод кластеризації групує веб-користувачів зі схожими інтересами.

В роботі [8] запропоновано об'єднання веб-користувачів на основі еволюції відвідування веб-сторінок. Виявлені закономірності змін інформаційних потреб веб-користувачів використовуються для їх угруповання. Згенеровані на основі історичних веб-сесій кластери Web-користувачів, можуть бути використані для персоналізованих веб-додатків: веб-реклами і веб-кешування.

Щоб отримати інформацію про інтереси користувачів на веб-сторінках в роботі [9] досліджується поведінка клієнта за допомогою вивчення записів веб-журналу. Час, проведений на веб-сторінці, і типи скоєних операцій показують ступінь зацікавленості веб-користувача. Досліджувані дані представляють собою журнали користувачів, зібрані за шість місяців.

Будь-який Інтернет-портал може постійно удосконалюватися, спираючись на інформаційну потребу користувача. Для збору і аналізу даних про користувачів в роботі [10] використовується метод ройового інтелекту, завдяки якому виявляються «подорожі» веб-користувачів з однаковими інтересами. результати кластерізації порівнюються з методами DBSCAN і K-means.

Завдяки самоорганізації, простоті і швидкодії спрощена модель нейронної мережі Кохонена пропонується використовуватися в інформаційно-пошуковій системі. Для її успішного застосування необхідно вирішити завдання формування змістовного образу документа і ідентифікацію кластера.

Описано методики та алгоритми, використовувані для управління великими наборами даних. Показана доцільність застосування самоорганізованих карт для аналізу даних великої розмірності.

Виклад основного матеріалу. Розглянемо найбільш популярні методи розбиття груп об'єктів на кластери, які застосовуються для вирішення завдання персоналізації в Web.

Алгоритм CURE. Виконує ієрархічну кластеризацію з використанням набору визначальних точок для приміщення об'єкта в кластер. Призначений для кластеризация дуже великих наборів числових даних. Ефективний для даних низької розмірності, працює тільки на числових даних.

Алгоритм MST. Алгоритм мінімального покриває дерева будує граф з N-1 ребер так, щоб вони з'єднували всі N точок і володіли мінімальною сумарною довжиною. До недоліків алгоритму відносяться: обмежена придатність. Алгоритм найбільш підходить для виділення кластерів типу згущення або стрічок.

Алгоритм CLOPE. Призначений для кластеризації величезних наборів категорійних даних. До переваг відносяться високі масштабованість і швидкість роботи і якість кластеризації. Відрізняється простотою програмної реалізації. При цьому він забезпечує більш високу продуктивність і кращу якість кластеризації в порівнянні з багатьма ієрархічними алгоритмами.

Аналіз алгоритмів показує, що з усіх розглянутих тільки CLOPE задовольняє необхідним вимогам. Розглянемо алгоритм CLOPE стосовно до задачі кластеризації пошукових профілів.

Нехай є база пошукових профілів D, що складається з множини пошукових профілів $\{t_1,t_2,\ldots,t_n\}$. Кожен профіль t_i є набір пошукових запитів $\{i_1,\ldots,i_m\}$. Множина кластерів $\{C_1,\ldots,C_k\}$ є розбиття множини $\{t_1,\ldots,t_n\}$, $\forall 1\le i,j\le k$. Кожен елемент C_i називається кластером, n, m, k – кількість профілів, кількість запитів в базі профілів і число кластерів відповідно.

Кожен кластер C має наступні характеристики: $D(C)$ – множину унікальних пошукових запитів; $\vartheta(i, C)$ – частоту входжень запиту i в кластер C;

$$W(C) = |D(C)|; \tag{1}$$

$$H(C) = \frac{S(C)}{W(C)}, \tag{2}$$

$$S(C) = \sum_{i\in D(C)} \quad \vartheta(i, C) = \sum_{t_i\in C} \quad |t_i|. \tag{3}$$

Очевидно, що чим більше значення H, тим більше «схожі»два профілі. Тому алгоритм повинен вибирати такі розбиття, які максимізують H. Для більш якісного розбиття замість $H(C)$ можна використовувати градієнт:

$$\frac{H(C)}{W(C)} = \frac{S(C)}{W(C)^2}. \quad (4)$$

Глобальна функція вартості має максимальне значення:

$$profit(C_i, r) \to max, \quad (5)$$

$$(1) profit(C_i, r) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} |C_i|}{\sum_{i=1}^{k} |C_i|}, \quad (6)$$

де $|C_i|$ кількість обєктів в *i*-тому кластері, *k* – кількість кластерів, *r* – позитивне дійсне число більше *1*.

Розглянемо реалізацію алгоритму. Нехай пошукові профілі зберігаються в таблиці бази даних. Для побудови початкового розбиття, що визначається функцією *Profit(C,r)* потрібен перший прохід по таблиці профілів. Після цього потрібна незначна (1-3) кількість додаткових сканувань таблиці для підвищення якості кластеризації і оптимізації функції вартості. Якщо в поточному проході по таблиці, змін не відбулося, то алгоритм припиняє свою роботу.

При побудові початкового розбиття з таблиці читається черговий профіль і створюється новий кластер (окрема таблиця) або поміщається в уже існуючий кластер, який дає максимум *Profit(C,r).*

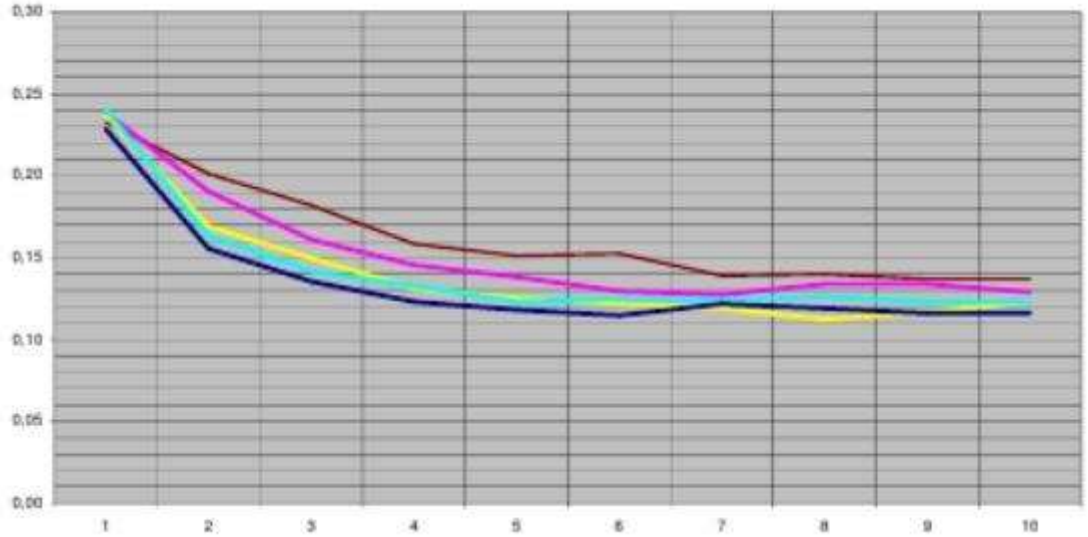

Рисунок 1 – Вероятність вибору по профілю користувача

На ітераційному етапі проглядається таблиця профілів і для кожного профілю вирішується завдання визначення кластера, якщо новий кластер максимізує *Profit(C, r)*, то профіль переноситься в цей кластер. На початку кожного циклу встановлюється індикатор переміщення *moved:=false*. Якщо в циклі відбувається переміщення профілю індикатор переміщення змінюється *moved:=true*. Ітерації завершуються, якщо значення *moved:= false* не зміниться. Після завершення ітерацій видаляються всі порожні кластери.

Алгоритм CLOPE є масштабованим, оскільки здатний працювати в обмеженому обсязі оперативної пам'яті комп'ютера. Під час роботи в оперативній пам'яті зберігається тільки поточна транзакція і невелика кількість інформації по кожному кластеру, яка

складається з: кількості транзакцій N, числа унікальних об'єктів (або ширини кластера) W, простий хеш-таблиці для розрахунку $\vartheta(i, C)$ і значення S площі кластера.

ВИСНОВКИ

Після кластеризації пошуковий профіль кінцевого користувача буде знаходитися у певному кластері.

Для пред'явлення користувачеві не переглянутих сторінок, відповідних постійної інформаційної потреби, проводиться розширення його пошукового профілю. Для цього пошукові запити, що входять до складу кластера C^*, ранжуються по частоті їх входження в кластер. У розширений пошуковий профіль вибирається деяка кількість l_m пошукових запитів з найбільшою частотою входження.

В результаті кластеризації встановлюються близькі пошукові профілі користувачів і на основі цього виявляються які раніше не переглянуті користувачем сторінки відповідні його постійним інформаційним потребам.

Список використаних джерел:

1. *Guha S., Rastogi R., Shim K.* «CURE: An Efficient Clustering Algorithm for Large Databases». / *S. Guha, R. Rastogi, K. Shim* // SIGMOD '98 Proceedings of the 1998 ACM SIGMOD international conference on Management of data, 1998, P. 73-84.

2. *Нильсен Я.* Веб-дизайн. Анализ удобства использования веб-сайтов по движению глаз / *Якоб Нильсен, Кара Перниче. – М. : Вильямс*, 2010. – 496 с.

3. *Нильсен Я.* Веб-дизайн. Web-дизайн: удобство использования веб-сайтов (юзабилити) / *Якоб Нильсен, Хоа Лоранжер. – М. : Вильямс*, 2007. – 368 с.

4. *Гарретт Д.* Элементы опыта взаимодействия / *Джесс Гарретт.* – Санкт-Петербург: Символ-Плюс, 2008. – 192 с.

5. *Нильсен Я.* Веб-дизайн. Дизайн Web-страниц. Анализ удобства и простоты использования 50 узлов / *Якоб Нильсен, Мари Тахир. – М. : Вильямс*, 2002. – 336 с.

6. *Hartigan J. A., Wong M. A.* "A K-means clustering algorithm,"/*J. A. Hartigan, M.A. Wong* //Applied Statistics, 1979, 28, 100 108.

7. *Xiao J.* Clustering of web users using session-based similarity measures/ *J. Xiao, Y. Zhang* // Computer Networks and Mobile Computing, 2001. Proceedings. 2001 International Conference on. – IEEE, 2001. – P. 223–228. DOI: 10.1109/ ICCNMC.2001.962600

8. *Chen L.* COWES: Web user clustering based on evolutionary web sessions / *L. Chen, S. S. Bhowmick, W. Nejdl* // Data & Knowledge Engineering. – 2009. – Vol. 68, No. 10. – P. 867–885. DOI:10.1016/j.datak.2009.05.002

9. *Selvakumar K.* Enhanced K-Means Clustering Algorithm for Evolving User Groups / *K. Selvakumar, L. S. Ramesh, A. Kannan* // Indian Journal of Science and Technology. – 2015. – Vol. 8, No. 24. – P. 1. DOI: 10.17485/ijst/2015/v8i24/80192

10. *Ganesan S.* Evolving interest based user groups using PSO algorithm / *S. Ganesan, A. I. U. Sivaneri, S. K. Selvaraju* // Recent Trends in Information Technology (ICRTIT), 2014 International Conference on. – IEEE, 2014. – P. 1–6. DOI: 10.1109/ICRTIT.2014.6996196

Левицькая Т.А., Чварков М.Д.

ИССЛЕДОВАНИЯ СОВРЕМЕННЫХ ТЕНДЕНЦИЙ ИСПОЛЬЗОВАНИЯ КЛАСТЕРИЗАЦИИ ПРИ РЕШЕНИИ ЗАДАЧИ ПЕРСОНАЛИЗАЦИЯ САЙТА

Данная работа посвящена решению проблем персонализации пользователей по их запросам. Анализ показал, что наилучшим методом кластеризации поисковых профилей, в данном случае, является метод кластеризации CLOPE. Разработан масштабируемый алгоритм, отличается большой масштабностью и легкой сложностью реализации

При кластеризации происходит анализ близких поисковых профилей пользователей, которые обращаются к системе, после чего ему отображаются ранее не просмотренные страницы, на основе поискового профиля пользователя. На основе разработанной модели и алгоритма предложена система персонализации, которая может интегрироваться с веб-сайтами для повышения эффективности доступа к релевантной для пользователя информации. Примером персонализации на основе текущих потребностей есть популярные в настоящее время рекомендательные системы, которые работают на основе поисковых профилей пользователей. Посетители, которые будут впервые оказываться на Web-сайте или веб-ресурсе любой организации, будут судить о ней по качеству и значимости опубликованных материалов.

Для крупных веб-ресурсов существует актуальная задача скорой навигационной и поисковой поддержки пользователей. Его можно решить путем персонализации контента, с учетом потребностей и особенностями поведения конечного пользователя. При персонализации веб-страниц будет осуществляться динамическое изменение содержимого веб-ресурса под конкретные нужды пользователя. В результате не только пользователь будет «общаться» с веб-страницей, но и сам сайт будет обращаться к любому, который попал на страницу, не как к части общей массы, а как до конкретного человека, что имеет свои личные интересы - персонально. Для решения вопроса кластеризации был выбран алгоритм CLOPE, которых предназначен для кластеризации больших объемов данных. В алгоритме CLOPE, во время работы сохраняется небольшое количество данных по каждому кластеру, при минимальном количестве сканирований. Целью данной статьи является публикация результатов исследования современных тенденций использования кластеризации при решении задачи персонализации.

Ключевые слова: *персонализация сайта, веб-ресурс, алгоритм, кластеризация, пользователь*

Levitskaya T., Chvarkov M.

RESEARCH OF MODERN TENDENCIES OF USE OF CLUSTERIZATION IN SOLUTION OF THE PROBLEM OF PERSONALIZATION OF THE SITE

This paper is devoted is devoted to solving the problems of personalization of users through their queries. The analysis showed that the best method of clustering of search profiles in this case is CLOPE clustering method. Developed a scalable algorithm, is of great scale and easy implementation complexity

Clustering is the analysis of the close search of profiles of users who access the system, then displays the previously viewed page based on the search profile of the user. Based on the developed model and the algorithm proposed personalization system that can integrate with web sites to improve the efficiency of access to relevant user information. An example of personalization based on current needs there is a currently popular recommendation systems that are based on search

user profiles. Visitors who will be first to be on the Web site or a web resource of any organization should be judged by the quality and importance of published materials.

For large web resources, there is an urgent task of emergency navigation and search user support. It can be solved by the personalization of content based on the needs and behaviors of the end user. When you personalize the web pages will be dynamically change the content of the web resource to the specific needs of the user. As a result, the user will "communicate" with the web page, but the site itself will appeal to anyone who got to the page, not as part of the total mass, and as to the particular person that has their personal interests, personally. To address the issue of clustering was chosen the algorithm of CLOPE, which is suitable for clustering large amounts of data. The algorithm CLOPE, during operation, is maintained a small amount of data for each cluster, with a minimum number of scans. The purpose of this article is to publish the results of the study of modern trends in the use of clustering in solving the problem of personalization.

Keywords: *site personalization, web resource, algorithm, clustering, user*
Рецензент: доцент, канд. техн. наук Міроненко Д.С.
Статья поступила

УДК 004.42

Міроненко Д. С.[1], Вонярха О.В.[2]

ДОСЛІДЖЕННЯ МОЖЛИВОСТЕЙ ВИКОРИСТАННЯ СУЧАСНИХ ТЕХНОЛОГІЙ ДЛЯ ПІДВИЩЕННЯ НАДІЙНОСТІ І ПРАЦЕЗДАТНОСТІ КОРПОРАТИВНОГО ПОШТОВОЇ СЕРВЕРА

На сьогоднішній день існує достатня кількість різних як платних, так і безкоштовних поштових сервісів, які надають прийнятний рівень зручності використання. У статті описані методи реалізації поштового сервісу на основі досліджень в області структурної організації пошти підприємства, а так само практичних випробувань поштових серверів, бізнес призначення, в середніх і великих IT-компаніях.

Наведено варіант підбору програмного-стека з актуальних на даний момент елементів відкритого програмного забезпечення, які виконують конкретну роль в роботі поштового сервера. Описані передові практики використання сучасних технологій захисту передачі конфіденційної інформації в момент передачі через глобальну мережу. А також актуальні методи валідації власної пошти, та її захисту від підміни третіми особами. Представлені основні методи захисту від спаму, і захисту користувачів від небажаної пошти. Дані методи і практики, можна використовувати як посібник для організації поштового сервера, або ж як рекомендації під час налаштування сервера з подібною специфікою.

Було встановлено, що реалізація зручного поштового сервісу на базі відкритого програмного забезпечення може бути виконано на основі програмного забезпечення Postfix спільно з Dovecot. Postfix повністю відповідає поставленим вимогам, з можливістю роботи в тандемі з іншими елементами стека, він є простим в подальшому адмініструванні. Dovecot – найефективніше рішення прийому електронної пошти у зв'язці з Postfix, в сфері

[1]*завідувач кафедри інформатики, доцент, кандидат технічних наук ДВНЗ «Приазовський державний технічний університет», м. Маріуполь, mironenko_ds@ukr.net*
[2]*бакалавр, ДВНЗ «Приазовський державний технічний університет», м. Маріуполь, happyvaran28@gmail.com*