КОМП'ЮТЕРНІ НАУКИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

УДК 004.89: 004.048 **DOI: 10.30977/BUL.2219-5548.2022.97.0.7**

ДОСЛІДЖЕННЯ КОГНІТИВНИХ СЕРВІСІВ ДЛЯ ПОШУКОВОЇ ОПТИМІЗАЦІЇ САЙТІВ

Бакланов О. М.[1], Безкоровайний В. В.[1], Колесник Л. В.[1]
[1]Харківський національний університет радіоелектроніки

Анотація. *Визначено завдання та здійснено експериментальне дослідження ефективності використання методів машинного навчання для побудови моделей автоматичної класифікації вебсторінок за ступенем адаптації до рекомендацій пошукової оптимізації SEO. Результати дослідження підходів та методів машинного навчання створюють засади для підвищення ефективності роботи пошукових систем. Вони можуть бути використані під час розроблення автоматизованого програмного забезпечення для підтримки роботи SEO в технологіях проведення аудиту для виявлення вебсторінок, які потребують оптимізації, та спамсторінок.*

Ключові слова: *ключове слово, когнітивний сервіс, машинне навчання, пошукова оптимізація, штучний інтелект.*

Вступ

В умовах конкуренції на ринку товарів і послуг, що постійно зростає, швидкими темпами розширюються сфери застосування електронних засобів реклами й надання рекомендацій з використанням технологій internet. Це висуває нові вимоги до зручності та доступності рекламних і пошукових сайтів як з боку користувачів мережі, так і щодо технологій пошукових роботів.

Сучасні пошукові системи стають все більш привабливими, зручними й ефективними. Одним з найбільш популярних підходів до вдосконалення сайтів є пошукова оптимізація (Search Engine Optimization – SEO). Її практичне використання дозволяє надавати пошуковим системам більше інформації, за результатами оцінки якої вміст сайтів індексується та відображається під час пошуків. Цьому сприяють когнітивні сервіси, які дозволяють відслідковувати попередню поведінку користувачів. Це обумовлює актуальність науково-прикладних завдань дослідження впливу застосування когнітивних сервісів для пошукової організації сайтів.

Аналіз публікацій

Сучасні пошукові системи з використанням сканування заповнюють бази даних або встановлюють індекси величезної кількості вебсторінок. У процесі сканування пошукові роботи завантажують вміст та посилання, знайдені на сторінках, на інші вебсторінки. Вони аналізують вміст вебсторінок, використовуючи алгоритми встановлення актуальності, якості, швидкості доступу, зручності для мобільних пристроїв тощо, що дозволяє визначити позицію або рейтинг результатів пошуку, які надаються користувачам [1].

Для підвищення якості пошукових процесів застосовують SEO-оптимізацію, яка є набором методів, призначених для поліпшення зовнішнього вигляду та позиціонування вебсторінок під час звичайного пошуку [2, 3]. Водночас аналізується SEO-контент, створений для залучення користувачів через пошукові системи. Найпопулярнішими форматами SEO-контенту є:

– сторінки продуктів;
– відео;
– інфографіка;
– блоги.

До найбільш значеннєвих факторів вебсторінок, які впливають на SEO-контент, належать:

– ключові слова;
– внутрішні посилання (гіперпосилання між двома сторінками на одному вебсайті);
– мета-заголовок та мета-описи [4].

Когнітивні сервіси, побудовані з використанням засобів штучного інтелекту, дозволяють оптимізувати контент сайтів, що задовольняє як видавців, так і користувачів. Такі сервіси використовуються для генерації заголовків, метатегів, начерків та контенту, виправлення речень, перекладання, генера-

ції/переписування абзаців з використанням ключових слів, аналізу змісту, розуміння мети користувача, семантичного контент-аналізу тощо.

Найбільш важливим є вплив використання ключових слів. З огляду на це одним з першочергових є завдання дослідження застосування когнітивних сервісів у SEO під час роботи з ключовими словами [5]. Для її оптимізації існують алгоритми класифікації, які можуть бути використані в SEO: дерево рішень (ДР) [6], Support Vector Machines (SVM) [7], Naive Bayes (NB) [8], k-найближчих сусідів (KNN) [9], логістична регресія (ЛР) [10].

Мета та постановка завдання

Метою дослідження є підвищення ефективності роботи пошукових систем завдяки встановлення та використання факторів, які мають найбільший вплив на ступінь SEO-оптимізації вебсторінок.

Для досягнення мети необхідно здійснити дослідження ефективності використання методів машинного навчання для побудови моделі класифікації вебсторінок за ступенем адаптації до рекомендацій щодо оптимізації SEO. Під час побудови моделі необхідно використовувати знання SEO-експертів.

Як фактори оптимізації вебсторінок використано такі елементи:

– текст на вебсторінці;

– текст у метатегах;

– посилання;

– зображення;

– код мови гіпертекстової розмітки HTML.

Необхідно визначити вплив цих факторів сторінки на ступінь оптимізації SEO за допомогою розроблених моделей класифікації.

Технологія дослідження

Дані для дослідження були вилучені вебсторінки каталога DMOZ. Як ключові слова використовувалися мітки його тек. Запропонована технологія дослідження передбачає чотири основні етапи (рис. 1):

– вибір випадкового вибору вебсторінок;

– класифікація вебсторінок за трьома попередньо визначеними експертами з SEO категоріями на сторінці щодо певного набору ключових слів;

– побудову та аналіз моделей класифікації;

– вилучення важливих факторів із моделей класифікації.

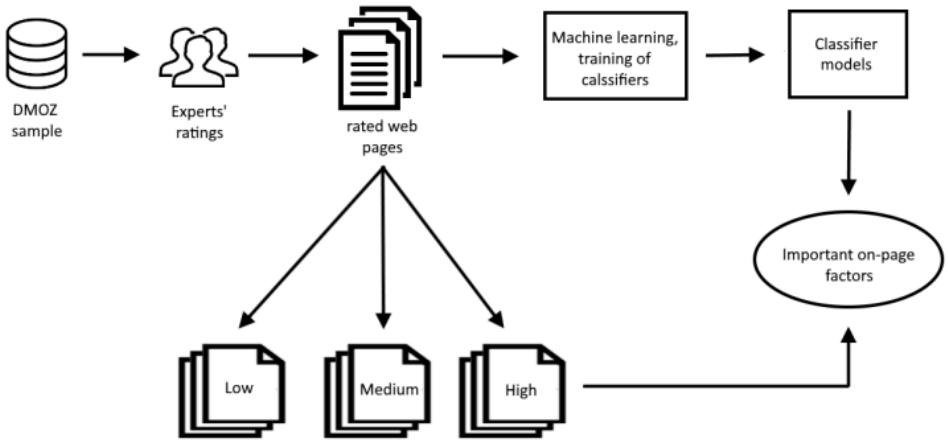

Рис. 1. Структурна схема технології дослідження когнітивних сервісів для пошукової оптимізації сайтів

На першому етапі в процесі випадкового відбору було визначено набір даних та вилучено ключові слова з назв категорій, до яких вони належать.

На другому етапі з використанням знань трьох незалежних експертів із SEO вебсторінки за заданими ключовими словами було розподілено на три попередньо визначені категорії якості: «Низька», «Середня» та «Висока». Отриманий у такий спосіб набір даних є вхідною інформацією для етапу побудови моделей класифікації з використанням алгоритмів машинного навчання.

Результати класифікації були використані для вилучення важливих факторів із моделей класифікації. Відповідні фактори, що впливають на класифікацію, були визначені та проаналізовані з вебсторінок, які мають най-

вищу якість SEO (належать до категорії якості «Висока»). Дослідження здійснювали з використанням вебсторінок, поданих англійською мовою.

Формування набору даних

Набір даних був сформований у процесі випадкового вибору 600 вебсторінок з каталога DMOZ, який часто використовується для пошуку інформації та класифікації вебсторінок, узагальнення їхніх текстів та визначення ключових слів. Він містить список вебсайтів в ієрархічних категоріях та підкатегоріях, залежно від теми вебсайту. Найвищі рівні в ньому мають такі категорії: «Мистецтво», «Бізнес», «Комп'ютери», «Ігри», «Здоров'я», «Дім», «Новини», «Відпочинок», «Довідковий», «Регіональний», «Наука», «Покупки», «Суспільство», «Спорт», «Каталог для дітей і підлітків» та «Світ» (табл. 1).

Таблиця 1 – Кількість вебсторінок в наборі категорій DMOZ

Категорія	Кількість сторінок	Категорія	Кількість сторінок
Мистецтво	63	Довідковий	31
Бізнес	69	Наука	39
Комп'ютери	58	Покупки	59
Ігри	35	Суспільство	31
Здоров'я	59	Спорт	42
Дім	43	Каталог для дітей та підлітків	34
Відпочинок	37	**Разом**	600

Рейтинг вебсторінок

Три незалежні експерти з SEO E_i, $i=\overline{1,3}$ відповідно до їхніх уподобань і досвіду на підставі заданих ключових слів і рекомендацій SEO поділили вебсторінки з набору даних на три попередньо визначені класи: «Низький SEO», «Середній SEO» і «Високий SEO» (табл. 2). Така початкова класифікація необхідна для етапу навчання класифікаторів.

Таблиця 2 – Результати експертного аналізу категорій вебсторінок

Клас	E_1	E_2	E_3	Цільовий
Низький SEO	180	119	112	146
Середній SEO	307	341	341	293
Високий SEO	113	140	147	161
Разом	600	600	600	600

Для перевірки узгодженості думок експертів була використана статистика Каппа [11]. Загальна середня оцінка статистики Каппа становить 0.4445, що зазвичай є достатньо позитивним результатом. Після цього було здійснено аналіз зваженої статистики Каппа для кожної пари експертів, який продемонстрував прийнятний ступінь узгодженості їхніх думок (табл. 3).

Таблиця 3 – Зважена статистика Каппа

Експерти	E_1	E_2	E_3
E_1	–	0.637	0.564
E_2	0.637	–	0.662
E_3	0.564	0.662	–

Створення списку незалежних змінних

Як незалежні змінні використано основні характеристики вебсторінки з повного списку (табл. 4). У подальшій роботі ці змінні використовуються для підготовки моделей класифікації.

Таблиця 4 – Список незалежних змінних

Група	Код змінної	Опис
Заголовок сторінки	tlen	Довжина контенту HTML-тегу «title» (лічильник слів)
Експерт	tkw	Частота ключового слова в HTML-тегу «title»
	mlen	Довжина контенту в HTML-тегу
	mkw	Частота ключового слова в HTML-тегу метаопису
Заголовки	h1	Кількість входжень HTML-тегу h1
	h1len	Середня довжина контенту HTML-тегу h1
	h1kw,	Частота ключового слова в HTML-тегу h1
	h2	Кількість входжень HTML-тегу h2
	h2klen	Середня довжина контенту HTML-тегу h2
	h2kw	Частота ключового слова у HTML-тезі h2
	h3	Кількість входжень HTML-тегу h3
	h3klen	Середня довжина контенту HTML-тегу h3
	h3kw	Частота ключового слова в HTML-тегу h3

Закінчення таблиці 4

Зображення	alt	Кількість входжень атрибута HTML alt в тег img (якщо alt містить вміст)
	altkw	Кількість входжень HTML-атрибута alt
Посилання	linkkw	Частота ключових слів в якорному тексті
	linkout	Кількість вихідних посилань
URL	urllen	Довжина URL (символів)
	urlkw	Частота ключових слів в URL
Текст	txtlen	Довжина тексту тіла сторінки
	txtkw	Частота ключових слів у тілі сторінки

Як залежну змінну було використано рейтинг експертів за кожною вебсторінкою, який визначив можливі значення класу: «1 – низький SEO», «2 – середній SEO» і «3 – високий SEO». Значення незалежних змінних зчитувалися автоматично за допомогою спеціально розробленого скрипту Python.

У процесі вилучення частот ключових слів використовувався стемер Портера [12]. Сформований набір даних містив 600 екземплярів, описаних 21 незалежною та 1 залежною змінними. Аналіз кореляції між змінними демонструє, що лише декілька пар змінних мали кореляцію, що є вище ніж 0.5: Mkw-Mlen, h1-h1len, h2-h2len, h3-h3len.

Вилучення факторів на сторінці

Аналіз набору даних може виявити, які конкретні фактори (незалежні змінні) можуть бути більш або менш важливими в прогнозуванні класу. Найпростіший спосіб досягти цього – ранжувати змінні відповідно до їхнього співвідношення зі змінною класу. Інший підхід полягає у використанні дерев рішень і властивостей прогнозування певної змінної.

Заходи змінної важливості, які використовуються в деревах рішень, можна використовувати для ранжування не лише тих змінних, які вибрано для розгалуження у вузлах дерева, а й усіх змінних, які були визначені. Водночас було використано коефіцієнт посилення інформації, який використовується в деревах рішень, щоб визначити, який атрибут необхідно розділити. Він розраховується як міра зменшення ентропії, отримана розщепленням за певним атрибутом [13].

Усереднюючи результати кількох дерев, побудованих на *n* вибірках і *m* предикторах (замість повного набору даних), можна зменшити дисперсію в методах на основі дерев і отримати кращі результати. Цей метод має назву «Випадковий ліс». Він може використовуватися для визначення важливих змінних.

Метод «Рельєф» для присвоєння ваги релевантності кожній ознаці використовує алгоритм «Найближчого сусіда». Поширеною мірою змінної важливості в задачах машинного навчання є Хі-квадрат [14]. Його використовують для перевірки зв'язку між предикторами (незалежними) та цільовою змінною (залежною). Змінні ранги, отримані за допомогою цих показників, наведені в табл. 5.

Змінними, для яких були отримані найвищі значення в більшості тестів (кореляція (К), приріст інформації (ПІ), хі-квадрат (Хі), рельєф (Р), випадковий ліс (ВЛ)), були такі: Tkw (частота ключових слів у тегу «title»), Mlen (довжина тегу «meta description»), Mkw (частота ключових слів у тегу «мета-опис»), h1Kw (частота ключових слів у тегу «H1») і txtKw (частота ключових слів у тілі сторінки). Тому ці змінні можна визначити як важливі фактори для оптимізації сторінок.

Таблиця 5 – Важливість змінних на основі різних тестів

Змінна	К	ПІ	Хі	Р	ВЛ
Tlen	0.25	0.05	0.32	2.26×10^{4}	21.8
Tkw	0.42	0.12	0.47	6.65×10^{3}	42.1
Mlen	0.38	0.20	0.56	3.15×10^{4}	51.1
Mkw	0.57	0.23	0.62	1.70×10^{4}	42.0
h1	0.02	0.03	0.27	6.64×10^{1}	10.3
h1len	0.03	0.03	0.26	-1.22×10^{2}	11.7
h1kw	0.30	0.06	0.36	7.46×10^{3}	15.4
h2	0.13	0.00	0.00	-8.69×10^{3}	15.8
h2len	0.09	0.00	0.00	-7.56×10^{3}	17.2
h2kw	0.27	0.05	0.30	1.96×10^{3}	13.1
h3	0.07	0.00	0.00	4.78×10^{3}	7.0
h3len	0.05	0.00	0.00	-4.67×10^{3}	7.6
h3kw	0.18	0.04	0.27	-5.10×10^{3}	4.5
alt	0.12	0.00	0.00	-3.28×10^{3}	8.7
altKw	0.24	0.04	0.29	-2.45×10^{2}	5.5
linkKw	0.27	0.05	0.33	2.80×10^{3}	12.3
linkOut	0.08	0.00	0.00	7.93×10^{1}	13.1
urlLen	0.01	0.00	0.00	6.23×10^{3}	1.6
urlKw	0.21	0.05	0.31	-7.25×10^{3}	11.3
txtLen	0.00	0.03	0.23	1.23×10^{3}	6.7
txtKw	0.26	0.09	0.42	9.74×10^{3}	31.8

Результати оцінювання моделі

Аналіз моделі здійснювався методом затримки та перехресної перевірки. Під час використання методу затримки дві третини

набору даних були використані для навчання моделі, а одна третина для її перевірки. Для перехресної валідації використовувався метод 10-кратної вкладеної перехресної перевірки (по 10 ітерацій у внутрішньому та зовнішньому циклах).

Для навчання моделей класифікації використовувався інструмент R з пакетом MLR. З використанням точності класифікації як міри оцінки для навчання використовувалося п'ять алгоритмів класифікації: дерево рішень (ДР), Support Vector Machine (SVM), Naive Bayes (NB), kNN та логістична регресія (ЛР). Усі гіперпараметри класифікатора були налаштовані за допомогою методу пошуку сітки. Для нормалізації даних було використано метод min-max, а для перехресної перевірки – метод стратифікації.

У табл. 6 наведено результати налаштування оптимальних значень гіперпараметрів (ОЗГ) й оцінки точності за методами затримки (ТМЗ) та перехресної перевірки (ТМПП).

Таблиця 6 – Оптимальні значення гіперпараметрів

Класифікатор	ОЗГ	ТМЗ	ТМПП
Дерево рішень	C = 0.49, M = 16	65.67 %	67.53 %
Naive Bayes	–	58.71 %	54.69 %
kNN	K = 45, p = 2	65.17 %	69.67 %
SVM	C = 7.74×103, σ = 0.000464	62.68 %	66.18 %
Логістична регресія	C = 2	62.19 %	62.99 %

Результати всіх п’яти класифікаторів порівняли з точністю базового рівня. У машинному навчанні базова модель є найпростішою. Як базу для порівняння використано алгоритм «ZeroR», його точність є орієнтиром для порівняння продуктивності інших алгоритмів машинного навчання.

У нашому випадку базовою була частка основного класу в наборі даних другого («Середнього SEO»). Мажоритарний клас спостерігався в 293 із 600 випадків, що становить 48,83 % набору даних. У 10-кратній перехресній перевірці середня точність усіх п’яти класифікаторів була вище ніж цей поріг (діаграми прямокутника на рис. 2). Червона лінія визначає точність базової (ТБЛ) – 48,83 %. Центральна лінія в квадратах позначає медіану, а точка – середнє. Медіана і середнє значення для всіх класифікаторів розташовані вище червоної лінії.

Моделі, побудовані для класифікації вебсторінок за рівнем їхньої SEO-оптимізації, мали більший рівень точності, ніж класифікація в мажоритарному класі. Щоб отримати статистично значеннєві висновки, необхідно здійснити остаточну перевірку за допомогою статистичних тестів.

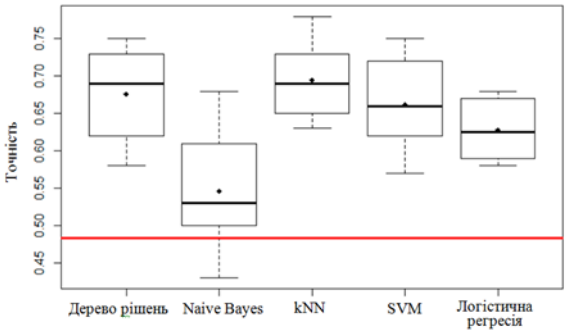

Рис. 2. Порівняння точності класифікаторів з базовою точністю

Для порівняння більшої кількості класифікаторів на одному наборі даних можна використовувати параметричні та непараметричні тести для залежних вибірок. Найпоширенішим параметричним тестом є t-критерій, який вимагає нормальності розподілу [15]. Оскільки в цьому дослідженні дотриматися цієї вимоги здебільшого було неможливо, були використані непараметричні тести, зокрема тести Макнемара та Вілкоксона.

Тест Макнемара використовується, якщо існує окремий набір даних для перевірки (відрізняється від навчального набору). Він не застосовується під час перехресної перевірки [16].

Таким чином, у цьому дослідженні можна використати тести Макнемара лише за допомогою методу тривалої перевірки. Цей тест вимагає побудови таблиць непередбачених обставин для кожної пари класифікаторів, тобто між точністю класифікації та точністю базового рівня. Результати наведені в табл. 7.

Таблиця 7 – Результати тесту Макнемара для порівняння алгоритмів класифікації з базовою лінією

Класифікатор	Матриця узгодження		Макнемар X^2	p-значення
Дерево рішень	35	68	9.9417	0.001616
	35	66		
SVM	35	68	6.75	0.009375
	40	61		

Закінчення таблиці 7

Логістична регресія	40	63	5.703	0.01694
	38	63		
kNN	44	59	5.6277	0.01768
	35	66		
NaïveBayes	36	67	0.28346	0.5944
	60	41		

Статистика тесту Макнемара X^2 має бути більше ніж 3.841 (розподіл X^2 з одним ступенем свободи та $\alpha = 0.05$), щоб можна було зробити висновок з 95-відсотковою впевненістю, що точність класифікації значно відрізняється від точності базового рівня. Тести демонструють, що так було з усіма перевіреними класифікаторами, крім «Наївного Байєса», який демонструє гірші результати.

Критерій знакового рангу Вілкоксона – це ще один непараметричний тест, який може визначити відмінності (або подібність) між розподілами двох вибірок з однієї сукупності. Він відомий як альтернатива параметричному t-критерію, оскільки не вимагає дотримання умови нормального розподілу [17]. У ньому спочатку обчислюються всі відмінності між відповідними значеннями змінних, які потім ранжуються від найменшого до найбільшого, ігноруючи полярність і підсумовуючи позитивні та негативні ранги (W^+ і W^-). Тоді значення Вілкоксона обчислюється за співвідношенням $W = min\{W^-, W^+\}$.

Якщо кількість спостережень $n \le 25$, то значення p вибирають в таблиці критичних значень Вілкоксона і порівнюють зі значенням W, щоб вирішити, чи необхідно відхиляти нульову гіпотезу [18]. Якщо $n > 25$, то розподіл вважається нормальним і можна розрахувати z-статистику.

З огляду на результати 10-кратної перехресної перевірки класифікаторів ($n = 10$) значення W було розраховано для кожної пари точності базової лінії класифікації (табл. 8).

Значення p демонструють, що всі перевірені класифікатори отримали значно кращі результати, ніж точність базового рівня. Дещо гірші результати були отримані для Naive Bayes, але вони все ж значно кращі, як порівняти з точністю базового рівня.

Для аналізу відмінностей між точністю класифікації побудованих моделей використано критерій Фрідмана. Це непараметричний тест, який відомий як альтернатива параметричному тесту ANOVA [19]. За допомогою критерію Фрідмана підтверджено наявність достовірної різниці в точності побудованих моделей класифікації на рівні значення $p = 0.05$. Для визначення того, між якими парами класифікаторів наявні суттєві відмінності, було використано додатковий постхостест Неменного (табл. 9).

Таблиця 8 – Результати тесту Вілкоксона (перехресної перевірки)

	ТБЛ	ДР	kNN	SVM	ЛР	NB
1	0.48	0.59	0.73	0.63	0.63	0.49
2		0.70	0.73	0.62	0.59	0.53
3		0.68	0.67	0.57	0.68	0.68
4		0.62	0.68	0.64	0.65	0.53
5		0.75	0.78	0.69	0.62	0.50
6		0.71	0.75	0.75	0.67	0.43
7		0.65	0.63	0.58	0.58	0.54
8		0.58	0.65	0.68	0.68	0.63
9		0.75	0.63	0.74	0.60	0.61
10		0.73	0.70	0.72	0.58	0.52
Середнє		0.68	0.69	0.66	0.63	0.55
W		55	55	55	55	52
p		0.001	0.001	0.001	0.001	0.009

Таблиця 9 – Порівняння відмінностей між парами моделей класифікації за критерієм Неменного

Клас	LH	NB	kNN	SVM
NB	0.0377	–	–	–
kNN	0.9692	0.0048	–	–
SVM	0.9986	0.0808	0.8896	–
ЛР	0.8267	0.3925	0.4357	0.9371

Відповідно до результатів можна зазначити, що існують значні відмінності між наївним Байєсовим алгоритмом і алгоритмом дерева рішень (для $p < 0.05$), а також між наївним Байєсом і KNN.

Висновки

Пошукова оптимізація вебсторінок передбачає процеси оптимізації змісту сторінки та отримання максимально якісних зворотних посилань. У цій статті досліджувався підхід машинного навчання для визначення ступеня адаптації вебсторінки до рекомендацій щодо SEO.

Класифікатори були вибрані з набору даних вебсторінок та розподілені трьома незалежними експертами з SEO на три попередньо визначені категорії: «Низький SEO», «Середній SEO» та «Високий SEO». Тестуючи п’ять основних класифікаторів (дерева

рішень, наївний Байєс, логістична регресія, KNN і SVM), моделі отримали більшу точність (від 54,69 % до 69,67 %), ніж точність базової лінії (48,83 %), яка є частиною більшості «Середнього SEO» класу.

За результатами дослідження було підтверджено, що за допомогою алгоритмів класифікації, побудованих на основі машинного навчання і знань експертів, можна здійснювати налаштування вебсторінок до рекомендацій SEO.

Для визначення факторів, релевантних для SEO, було використано алгоритм дерева рішень. Набір даних, сформований у цьому дослідженні, може бути використаний в подальших роботах SEO-факторів рейтингу сторінки. Методи, використані в цьому дослідженні, не були специфічними для пошукової системи та мови. Розглянуті методи можуть бути адаптовані до різноманітних пошукових систем та застосовані до різних мов за умови, що для них розроблено алгоритм для стемпінгу або лемматизації.

Для підвищення точності результатів доцільним є збільшення кількості експертів із SEO та (або) цільових класів у процесі маркування сторінок.

Результати дослідження можуть бути використані під час розроблення автоматизованого програмного забезпечення для підтримки роботи SEO в технологіях здійснення аудиту для визначення вебсторінок, які потребують оптимізації, і в процесах визначення спамсторінок.

Набір даних, сформований у цій статті, може бути використаний в подальших дослідженнях факторів SEO або методів класифікації вебсторінок. Напрямами подальшого розвитку отриманих результатів може бути дослідження інших методів машинного навчання або використання інших позасторінкових факторів.

Література

1. Shenoy A., Prabhu A. Introducing SEO. Dordrecht: Springer Nature, 2016. 147 p.
2. Enge E., Spencer C., Stricchiola J. The art of SEO. Mastering Search Engine Optimization. Sebastopol: O'Reilly Media, 2022. 149 p.
3. Shane D. SEO decoded. 39 search engine optimization strategies to rank your website for the toughest of keywords. North Charleston: CreateSpace, 2016. 210 p.
4. Колесник Л. В., Бакланов О. М. Актуальність задачі SEO-оптимізації сайту. Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління: тези доповідей XX Міжнародної науково-технічної конференції. Баку – Харків – Жиліна. 2022. Т. 2. С. 52.
5. Колесник Л. В., Бакланов О. М. Аналіз способів застосування когнітивних сервісів для SEO-оптимізації сайту. Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління: тези доповідей XX Міжнародної науково-технічної конференції. Баку – Харків – Жиліна, 2022. Т. 2. С. 53.
6. Lantz B. Machine Learning with R. Birmingham – Mumbai: Packt Publishing, 2013. 396 p.
7. Ben-Hur A., Weston J. A Users Guide to Support Vector Machines. Data Mining Techniques for the Life Sciences. Methods in Molecular Biology. 2010. 609 p.
8. Hand D. J., Yu K. Idiot's Bayes – not so stupid after all? International Statistical Review. 2001. No 69 (3). P. 385–399.
9. Antonov A. A. From Artificial Intelligence to Human Super-Intelligence. Computer Information Systems. 2011. Vol. 2. No 6. P. 1–6.
10. Gupta S., Aggarwal A. Study of Search Engine Optimization. Research in Engineering & Applied Sciences. 2012. Vol. 2. No. 2. P. 1529–1536.
11. Дрейпер Н, Смит Г. Прикладной регрессионный анализ. Множественная регрессия = Applied Regression Analysis. Москва: Диалектика, 2007. 912 с.
12. Бакаев И. И., Шафиев Т. Р. Методы построения алгоритмов стемминга для массового языка. Проблемы вычислительной и прикладной математики. 2020. № 3(27). С. 146–154.
13. Aul V. Harnessing Search Engine Optimization Experience to Enhance the Visibility of Websites: Ph.D / University of West London. London, 2018. 485 p.
14. Hashemi M. Web page classification: A survey of perspectives, gaps, and future directions. Multimedia. Tools Apple. 2020. No 79. P. 11921–11945.
15. Kendall M. G. A New Measure of Rank Correlation. Biometrika. 2008. No 30. P. 81–93.
16. Abdullah K. D. Search Engine Optimization Techniques by Google's Top Ranking Factors: Website Ranking Signals. North Charleston: Independently published, 2017. 106 p.
17. Andersson V., Lindgren D. Ranking Factors to Increase Your Position on the Search Engine Result Page. Karlskrona: Theoretical and Practical Examples. 2017. 389 p.
18. Mavridis T., Symeonidis A. L. Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms. Engineering Applications of Artificial Intelligence. 2015. No 41. P. 75–91.
19. On-Page Search Engine Optimization: Study of Factors Affecting Online Purchase Decisions of Consumers // Sujata, J., Noopur, S., Neethi, N.,

Jubin, P., Udit P. (2016). *Indian Journal of Science and Technology*. No 9. p. 1–10.

References

1. Shenoy A., Prabhu A. Introducing SEO. Dordrecht: Springer Nature, 2016. 147 p.
2. Enge E., Spencer C., Stricchiola J. The art of SEO. Mastering Search Engine Optimization. Sebastopol: O'Reilly Media, 2022. 149 p.
3. Shane D. SEO decoded. 39 search engine optimization strategies to rank your website for the toughest of keywords. North Charleston: CreateSpace, 2016. 210 p.
4. Kolesnyk, L. V., Baklanov, O. M. (2022). The relevance of the task of SEO-optimization of the site. *Modern directions of the development of information and communication technologies and management tools: theses of reports of the twelfth international scientific and technical conference*. Baku – Kharkiv – Zhilina. Vol. 2. P. 52. [in Ukraine]
5. Kolesnyk, L. V., Baklanov, O. M. (2022). Analysis of ways to use cognitive services for SEO site optimization. *Modern directions of the development of information and communication technologies and management tools: theses of reports of the twelfth international scientific and technical conference*. Baku – Kharkiv – Zhilina. Vol. 2. P. 53. [in Ukraine]
6. Lantz B. Machine Learning with R. Birmingham – Mumbai: Packt Publishing, 2013. 396 p.
7. Ben-Hur A., Weston J. A Users Guide to Support Vector Machines. Data Mining Techniques for the Life Sciences. Methods in Molecular Biology. 2010. 609 p.
8. Hand D. J., Yu K. Idiot's Bayes – not so stupid after all? International Statistical Review. 2001. No 69 (3). P. 385–399.
9. Antonov A. A. From Artificial Intelligence to Human Super-Intelligence. Computer Information Systems. 2011. Vol. 2. No 6. P. 1–6.
10. Gupta S., Aggarwal A. Study of Search Engine Optimization. Research in Engineering & Applied Sciences. 2012. Vol. 2. No. 2. P. 1529–1536.
11. Draper, N, Smith, G. (2007). *Applied regression analysis. Multiple regression = Applied Regression Analysis*. Moskov: Dialectics. 912 p. [in Russian]
12. Bakaev, I. I., Shafiev, T. R. (2020). Methods of constructing stemming algorithms for a mass language. *Problems of computational and applied mathematics*. No. 3(27). P. 146–154. [in Russian]
13. Aul V. Harnessing Search Engine Optimization Experience to Enhance the Visibility of Websites: Ph.D / University of West London. London, 2018. 485 p.
14. Hashemi, M. (2020). Web page classification: A survey of perspectives, gaps, and future directions. *Multimedia. Tools Apple*. No 79. P. 11921–11945.
15. Kendall M. G. A New Measure of Rank Correlation. Biometrika. 2008. No 30. P. 81–93.
16. Abdullah K. D. Search Engine Optimization Techniques by Google's Top Ranking Factors: Website Ranking Signals. North Charleston: Independently published, 2017. 106 p.
17. Andersson V., Lindgren D. Ranking Factors to Increase Your Position on the Search Engine Result Page. Karlskrona: Theoretical and Practical Examples. 2017. 389 p.
18. Mavridis T., Symeonidis A. L. Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms. Engineering Applications of Artificial Intelligence. 2015. No 41. P. 75–91.
19. On-Page Search Engine Optimization: Study of Factors Affecting Online Purchase Decisions of Consumers // Sujata, J., Noopur, S., Neethi, N., Jubin, P., Udit P. (2016). *Indian Journal of Science and Technology*. No 9. p. 1–10.

Бакланов Олексій Миколайович[1], магістрант кафедри системотехніки, тел.: +38 097-153-01-44, oleksii.baklanov@nure.ua,
Безкоровайний Володимир Валентинович[1], д.т.н., проф., професор кафедри системотехніки, тел.: +38 050-983-03-29, vladimir.beskorovainyi@nure.ua,
Колесник Людмила Володимирівна[1], к.т.н., доц., доцент кафедри системотехніки, тел.: +38 095-122-74-75, liudmyla.kolesnyk@nure.ua,
[1]Харківський національний університет радіоелектроніки, пр. Науки, 14, м. Харків, 61166, Україна.

Studying cognitive services for websites search engine optimization

Annotation. The subject *of research in the article is machine learning models for classifying web-pages by quality and compliance with SEO rules.* ***The goal*** *of the article is improving the efficiency of search engines by establishing and using factors that have the greatest impact on the degree of SEO optimization of web pages. The article solves the following* ***tasks:*** *study of the effectiveness of using machine learning methods to build a classification model that automatically classifies web pages according to the degree of adaptation to SEO optimization recommendations; assessment of the influence of relevant page factors (text on a web page, text in meta tags, links, image, HTML code) on the degree of SEO optimization using the developed classification models.* ***The following methods*** *are used: machine learning methods, classification methods and statistical methods.* ***The following results*** *were obtained: analysis of the effectiveness of the application of machine learning methods to determine the degree of adaptation of a web page to SEO recommendations was carried out; classifiers were trained on a data set of web pages randomly selected from the DMOZ catalog and rated by three independent SEO experts in*

the categories: "low SEO", "medium SEO" and "high SEO"; five main classifiers were tested (decision trees, naive Bayes, logistic regression, KNN and SVM), on the basis of which it was revealed that all the studied models received greater accuracy (from 54.69 % to 69.67 %) than the accuracy of the baseline (48.83 %); the results of the experiments confirm the hypothesis about the effectiveness of adapting web pages to SEO recommendations using classification algorithms based on machine learning. ***Conclusions.*** *It was confirmed that with the help of classification algorithms built on the basis of machine learning and the knowledge of experts, it is possible to effectively adjust web pages to SEO recommendations. The considered methods can be adapted for various search engines and applicable to different languages, provided that a stamping or lemmatization algorithm has been developed for them. The results of the study can be used in the development of automated software to support the work of SEO in audit technologies to identify web pages in need of optimization and in spam detection processes.*

Keywords: *keyword, cognitive service, machine learning, search engine optimization, artificial intelligence.*

Oleksii Baklanov[1], Master's Degree Candidate of the Department of System Engineering,
tel.: +38 097-153-01-44, oleksii.baklanov@nure.ua,
Volodymyr Bezkorovainyi[1]**,** Doctor of Engineering Sciences, Professor, tel.: +38 050-983-03-29,
vladimir.beskorovainyi@nure.ua,
Liudmyla Kolesnyk[1], Phd, Associate Professor,
tel.: +38 095-122-74-75,
liudmyla.kolesnyk@nure.ua,
[1]Kharkiv National University of Radio Electronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine.