

УДК 519.711.3

А.В. Цуканов, профессор, д-р техн. наук,**М.В. Потанина, ст. преподаватель***Севастопольский национальный технический университет**Ул. Университетская 33, г. Севастополь, Украина, 99053**E-mail: mem@sevgtu.sebastopol.ua***ПОСТРОЕНИЕ ПРОГНОЗИРУЮЩЕЙ МОДЕЛИ ДЛЯ ОЦЕНКИ РЕЛЕВАНТНОСТИ ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ ОРГАНИЗАЦИИ САЙТА**

Рассмотрена методика построения прогнозирующей модели для оценки релевантности информационной технологии организации сайта. Целью моделирования является получение модели с наименьшей ошибкой прогноза в заданной области возможных значений параметров сайта. Предлагается использовать модифицированный метод селекции регрессионных моделей.

Ключевые слова: *прогнозирующая модель, релевантность информационной технологии организации сайта, ошибка прогноза, селекция регрессионных моделей.*

В настоящий момент объем информации в Интернете, поддерживаемый поисковыми системами (Google, Яндекс и пр.) насчитывает десятки миллиардов web-страниц. При наличии такого объема информации важно, чтобы содержимое сайта доходило до своей целевой аудитории и не терялось среди моря ненужной информации.

В Интернете постоянно происходит обновление накопленной информации. Это приводит к необходимости совершенствования сайтов их владельцами. В связи с этим, важное значение принимает выбор информационной технологии, которая используется для создания, управления и обработки данных сайта. Приходится также решать такие задачи, как взаимодействие сайта с пользователем, поиск информации в базе данных, обеспечение доступности материала на различных платформах. Весь этот процесс совершенствования сайта требует проведения с ним постоянных экспериментов и сбора статистики по качеству процесса поиска информации, содержащейся на сайте.

Проблема совершенствования информационной технологии организации сайтов с целью повышения его релевантности постоянно находится в поле зрения авторов отечественной и зарубежной литературы.

Детальный анализ задач правильного пользования поисковыми системами, поднятия рейтинга новых неизвестных сайтов содержат публикации Колесниченко Д.Н. [1] и Шэри Тетроу [2].

В работах автора Щербакова А.Ю. [3] подробно рассказывается о существующих поисковых машинах и интернет-ресурсах, которые являются инструментами для аналитической работы, управления бизнесом и обеспечения его безопасности. Интернет-аналитика заключается в измерении, сборе, анализе, представлении и интерпретации информации о посетителях веб-сайтов с целью их улучшения и оптимизации. Основной задачей веб-аналитики является мониторинг работы веб-сайтов, на основании которого определяется веб-аудитория и изучается поведение веб-посетителей для принятия решений по развитию и расширению функциональных возможностей веб-ресурса.

Из подробного анализа публикаций по данной тематике видно, что большинство исследователей подчеркивают огромную важность Интернет-аналитики работы сайта. На основании этого делаются выводы об эффективности того или иного сайта. Однако, несмотря на это, не существует общепринятой методики построения прогнозирующей модели для оценки релевантности информационной технологии организации сайта.

Целью работы является построение прогнозирующей модели для оценки релевантности информационной технологии организации сайта возможным запросам пользователей на основе модифицированного метода селекции регрессионных моделей. При этом под целью моделирования предполагается получение модели с наименьшей ошибкой прогноза качества сайта в заданной области возможных значений его параметров.

Качество сайта можно оценить по тому, как пользователи Интернета находят этот сайт в случае поиска соответствующей информации, то есть как информация, представленная на сайте, соответствует запросам пользователей. Любой сайт в Интернете можно описать множеством параметров, таких как: объем сайта, количество веб-страниц сайта, периодичность обновления информации на сайте и др. Релевантность сайта запросам пользователей можно оценить по месту, которое сайт занимает в списке сайтов, полученных с помощью поисковых систем типа Google и Яндекс. Для того, чтобы оценить влияние параметров сайта на качество поиска хранящейся в нем информации, необходимо собрать статистику по влиянию этих параметров для сайтов аналогичной тематики.

Пусть матрица X — матрица выборочных значений k информационных параметров N сайтов, проанализированных и оцененных экспертами; y — релевантность сайта запросу, оцениваемая по месту

сайта в списке результатов поиска, Y — вектор столбец значений функции отклика y для всех N сайтов. Например, если имеется три параметра и данные по пяти сайтам, то матрицы X и Y могут иметь следующий вид:

$$X = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{bmatrix}; \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}.$$

Требуется определить структуру и параметры регрессионной модели для оценки результата поиска в зависимости от характеристик сайта

$$y = \vec{f}_j(\vec{x})^T \vec{a}_j + \varepsilon, \quad (1)$$

где $\vec{x}$ — вектор значений характеристик сайта, $\vec{f}_j$ — вектор известных функций для модели j , $\vec{a}_j$ — вектор неизвестных параметров модели, ε — случайная составляющая модели.

Анализ статистических данных по характеристикам сайтов показывает, что между этими характеристиками существует сильная корреляционная зависимость. Поэтому применение обычного метода линейного регрессионного анализа для таких данных ведет к значительным ошибкам.

Предлагается использовать следующий метод поиска модели, обладающей наибольшей прогнозирующей способностью.

На первом этапе определяется класс вложенных моделей.

В качестве критерия селекции модели предлагается использовать классический критерий Ср Маллоуза (Mallows) [5]

$$C_p = C_j / \hat{\sigma}^2 - N, \quad (2)$$

где $C_j = RSS_j + \gamma k_j \hat{\sigma}^2$; $\hat{\sigma}^2$ — оценка дисперсии ошибки случайной составляющей σ^2 ; полученная для модели наиболее высокого порядка; k_j — число неизвестных параметров в j -ой модели; γ — регулируемый коэффициент, который обычно принимается равным двум.

Так как матрица X имеет сильно коррелированные столбцы, то для оценки параметров регрессионных моделей предлагается использовать ридж-оценивание [4]

$$\hat{a}_{j(RIDG)} = (F_j^T F_j + r I_j)^{-1} F_j^T Y, \quad (3)$$

где $F_j^T = [f_j(x_1), f_j(x_2), \dots, f_j(x_N)]$ — матрица значений вектора функций f_j в N экспериментальных точках для j -ой модели.

Предполагается, что матрица F_j имеет полный ранг. Модель максимального размера с номером q будем считать истинной моделью.

Для каждой j -ой модели из вложенного класса находится оптимальный параметр регуляризации r . Так как этот параметр зависит от истинных значений коэффициентов уравнения регрессии, то оптимальное значение параметра регуляризации находится с помощью итерационной процедуры, которая состоит в следующем.

Определяются границы возможных значений параметров модели.

Для возможных диапазонов изменения отношения дисперсии случайной составляющей к максимальному значению функции отклика методом Монте-Карло находится зависимость ошибки прогноза в данной области от параметров [6]

$$L(q) = E \int_{w_1} f(x)^T (\alpha_{IST} - \hat{a}(r)) (\alpha_{IST} - \hat{a}(r))^T f(x) dx / \int_{w_q} dx = \sigma^2 sp F B J B^T + \alpha^T (I_q - F^T F B) J (I_q - B F^T F) \alpha, \quad (4)$$

где $B = (F^T F + r I_q)^{-1}$ и $J = \int_{w_1} f(x) f(x)^T dx / \int_{w_1} dx$; α — вектор истинных значений параметров.

Формула (4) состоит из двух частей. Первая, содержащая истинные коэффициенты $\gamma_1(r) = \alpha^T (I_q - F^T F B) J (I_q - B F^T F) \alpha$, показывает влияние смещения оценок, а вторая часть $\gamma_2(r) = \sigma^2 sp F B J B^T$ содержит дисперсию ошибки.

Определяются предварительные оценки коэффициентов уравнения регрессии обычным методом наименьших квадратов

$$\hat{\alpha}_{j(MNK)} = (F_j^T F_j)^{-1} F_j^T Y, \quad (5)$$

где $F_j^T = [f_j(x_1), f_j(x_2), \dots, f_j(x_N)]$ — матрица значений вектора f_j в N экспериментальных точках.

Оценка дисперсии случайной составляющей определяется из модели q максимального размера, которая считается истинной моделью

$$Y = F_q \alpha_q + \varepsilon. \quad (6)$$

Находятся оптимальные параметры регуляризации для ридж-регрессии каждого класса модели.

Находятся коэффициенты с помощью ридж-регрессии для каждой j -ой модели в соответствие с параметром регуляризации.

Выбирается наилучшая модель по критерию C_p – Маллоуза.

Для проверки работоспособности метода в качестве примеров рассматривались сайты по программному обеспечению в области современных методов управления предприятиями.

Для анализа было выбрано 20 сайтов: www.1c.ru, www.sap.com, www.intalev.ru, www.pentaho.com, www.tibco.com, www.adaptiveplanning.com, www.redhat.com, www.microsoft.com, www.alfresco.com, sourceforge.net, www.jaspersoft.com, www.mysql.ru, forge.ow2.org, www.ibm.com, www.wfmc.org, bpms.intalio.com, www.boc-eu.com, www.bpmg.org, www.softwareag.com, www.arenasimulation.com

Запросы в поисковой системе Google состояли из ключевых слов, связанных с понятием «информационные технологии по оценке эффективности бизнес-процессов».

Функция отклика y определялась как место сайта в списке ответов системы, которое интерпретировалось как релевантность структуры соответствующего сайта и запроса на поиск информационной технологии.

При этом был осуществлен поиск по 5 различным запросам на данную тему. В результате была получена матрица экспериментальных данных, состоящая из 100 строк и 10 столбцов. Первые девять столбцов матрицы соответствовали девяти параметрам сайта: x_1 — объем сайта; x_2 — общее число веб-страниц сайта; x_3 — число картинок (рисунков) на сайте; x_4 — возраст сайта; x_5 — периодичность обновления информации на сайте; x_6 — последнее обновление страниц сайта; x_7 — наличие замещающих надписей на рисунках (картинках); x_8 — размер шрифта, которым оформлены ключевые слова; x_9 — общее число гиперссылок сайта; последний столбец — функция отклика y .

Многие данные по оценке сайтов носят экспертный характер. При этом в оценивании сайтов принимали участие студенты Севастопольского национального технического университета.

Был проведен корреляционный анализ исходных данных по параметрам сайтов. Полученная матрица парных коэффициентов корреляции Пирсона показана на рисунке 1.

	x1	x2	x3	x4	x5	x6	x7	x8
x2	0.989							
x3	0.916	0.887						
x4	0.751	0.709	0.772					
x5	-0.575	-0.590	-0.452	-0.403				
x6	-0.497	-0.481	-0.399	-0.245	0.776			
x7	0.362	0.379	0.378	0.271	-0.375	-0.203		
x8	0.003	-0.023	-0.029	0.181	-0.250	-0.360	-0.028	
x9	0.912	0.899	0.863	0.717	-0.510	-0.378	0.395	0.050

Рисунок 1 — Корреляционная матрица параметров сайтов

Результаты корреляционного анализа показывают, что между многими параметрами существует сильная корреляционная зависимость. Например, между параметрами x_1 и x_2 коэффициент корреляции равен 0,989, что соответствует почти функциональной зависимости.

На рисунке 2 показаны результаты расчета уравнений регрессии обычным методом наименьших квадратов для полной модели.

Результаты линейного регрессионного анализа показывают, что большинство коэффициентов регрессионного уравнения незначимы и использовать полученную модель для прогноза нельзя.

В результате применения модифицированного метода селекции регрессионных моделей [6] была получена следующая прогнозирующая модель для оценки релевантности сайта

$$y = 15.5533 - 0.0235 x_1 + 0.3752 x_3. \quad (7)$$

Значение параметра регуляризации для этой модели $r_{opt} = 0,32$. Так как оценки коэффициентов модели являются смещенными, то применить стандартные критерии для оценки значимости

коэффициентов и самой модели нельзя. На рисунке 3 показаны результаты регрессионного анализа для модели аналогичной структуры, но оцененной обычным методом наименьших квадратов.

Predictor	Coef	StDev	T	P
Constant	17.707	6.192	2.86	0.005
x1	-0.03029	0.01469	-2.06	0.042
x2	0.0950	0.2544	0.37	0.710
x3	0.3696	0.1151	3.21	0.002
x4	-0.0021	0.2709	-0.01	0.994
x5	0.4073	0.5429	0.75	0.455
x6	-0.7422	0.7934	-0.94	0.352
x7	-0.082	1.094	-0.07	0.941
x8	-0.1108	0.3980	-0.28	0.781
x9	0.02415	0.04261	0.57	0.572

S = 4.623 R-Sq = 42.2% R-Sq(adj) = 36.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	9	1401.58	155.73	7.29	0.000
Error	90	1923.42	21.37		
Total	99	3325.00			

Рисунок 2 – Результаты построения модели методом наименьших квадратов

Predictor	Coef	StDev	T	P
Constant	15.832	1.074	14.75	0.000
x1	-0.023582	0.003660	-6.44	0.000
x3	0.36835	0.09625	3.83	0.000

S = 4.485 R-Sq = 41.3% R-Sq(adj) = 40.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1373.98	686.99	34.16	0.000
Error	97	1951.02	20.11		
Total	99	3325.00			

Рисунок 3 – Результаты регрессионного анализа для выбранной структуры модели

Так как модель для минимальной ошибки прогноза подбиралась с использованием модифицированного метода, то она имеет большую прогнозирующую способность, чем модель, полученная с помощью метода МНК.

Предложенный метод показал свою работоспособность и эффективность для анализа качества сайтов. В дальнейших исследованиях необходимо расширить спектр возможных запросов в поисковых системах и усовершенствовать метод для оценки значимости получаемых результатов, например, с использованием метода «бутстреп».

Библиографический список использованной литературы

1. Колесниченко Д.Н. Поисковые системы и продвижение сайтов в Интернете / Д.Н. Колесниченко. — М.: Вильямс, 2007. — 272 с.
2. Тероу Ш. Видимость в Интернете. Поисковая оптимизация сайтов / Ш. Тероу. — С-Пб: Символ-Плюс, 2009. — 288 с.
3. Щербаков А.Ю. Интернет-аналитика. Поиск и оценка информации в web-ресурсах. Практическое пособие / А.Ю. Щербаков. — М.: Книжный мир, 2011. — 80 с.
4. Hoerl A. Ridge regression: biased estimation for no orthogonal problems / A. Hoerl, R. Kennard // *Technometrics*. — 1970. — Vol. 12. — P. 55–67.
5. Айвазян С.А. Прикладная статистика: Исследование зависимостей: справ. изд./ С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин; под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1985. — 487 с.
6. Потанина М.В. Оценка качества идентификации сложного объекта управления в условиях мультиколлинеарности / М.В. Потанина // *Вестник СевНТУ. Сер. Автоматизация процессов управления: сб. науч. тр.* — Севастополь, 2009. — Вып. 95. — С. 135–140.

Поступила в редакцию 26.04.2012 г.

Цуканов О.В., Потаніна М.В. Побудова прогнозованої моделі для оцінки релевантності інформаційної технології організації сайту

Розглянуто методику побудови прогнозованої моделі для оцінки релевантності інформаційної технології організації сайту. Метою моделювання є одержання моделі, що буде мати найменшу помилку прогнозу в заданій зоні можливих значень параметрів сайту. Пропонується використати модифікований метод селекції регресійних моделей з використанням рідж-регресії та критерію C_p Маллоуза.

Ключові слова: прогнозована модель, релевантність інформаційної технології організації сайту, помилка прогнозу, селекція регресійних моделей.

Tsukanov A.V., Potanina M.V. Forecasting Model Construction for the Estimation of the Site Information Technology Relevance

The method of forecasting model construction for the estimation of the site information technology relevance is considered. A design purpose is a receipt of model which will have the least error of prediction in the set area of possible values of parameters of site. It is suggested to use the modified method of selection of regression models with ridge-regression and C_p criterion of Mallows.

Keywords: forecasting model, the site information technology relevance, the least error of prediction, the selection of regression models.