УДК 004.738.52

ПОСТРОЕНИЕ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ РАНЖИРОВАНИЯ РЕЗУЛЬТАТОВ РАБОТЫ ПОИСКОВЫХ СИСТЕМ GOOGLE И ЯНДЕКС

В.В. Зосимов[1], В.С. Степашко[2], А.С. Булгакова[1]
[1] *Николаевский национальный университет им. В.О.Сухомлинского, 54000 Николаев, ул. Никольская 24*
[2] *Международный научно-учебный центр информационных технологий и систем, 03680 Киев, пр. Академика Глушкова, 40*

zosimovvv@bk.ru, stepashko@irtc.org.ua , sashabulgakova1@gmail.com

Побудовано моделі ранжування результатів видачі пошукових систем Google і Яндекс із застосуванням індуктивних алгоритмів. Проведено порівняльний аналіз побудованих моделей та виявлено найбільш вагомі ознаки ранжування і характерні особливості моделей.
Ключові слова: *ранжування, пошукові системи, індуктивне моделювання, ітераційні алгоритми, МГУА.*

Models of ranking the return results of search engines Google and Yandex were built using inductive algorithms. Comparative analysis of constructed models was done and the most significant ranking features and models specific peculiarities found out.
Keywords: *ranking, search engines, inductive modeling, iterative algorithms, GMDH*

Построены модели ранжирования результатов выдачи поисковых систем Google и Яндекс с применением индуктивных алгоритмов. Проведен сравнительный анализ построенных моделей и выявлены наиболее весомые признаки ранжирования и характерные особенности моделей.
Ключевые слова: *ранжирование, поисковые системы, индуктивное моделирование, итерационные алгоритмы, МГУА*

Введение

В данной статье описаны результаты построения и сравнительного анализа моделей ранжирования результатов работы поисковых систем Google и Яндекс. Целью исследования является выявление характерных особенностей моделей ранжирования для дальнейшего поиска возможных путей их усовершенствования или построения более эффективных моделей.

В основе исследования лежит построение моделей ранжирования с применением обобщенного итерационного алгоритма МГУА на основе обучающей выборки, в качестве которой использовались результаты ранжирования поисковых систем.

На основе полученных в ходе экспериментов данных исследуется важность признаков ранжирования, а также сходство и различие моделей поисковых систем Google и Яндекс.

1. Построение модели ранжирования поисковой выдачи Google

В данной задаче моделируем процесс ранжирования веб-ресурсов поисковой выдачи системы Google (google.com.ua) для поисковой фразы «веб-программирование».

Для эксперимента было отобрано первых 50 сайтов поисковой выдачи по данному ключевому запросу. Матрица исходных данных X содержит 42 признака-фактора, которые численно характеризуют каждый сайт (см. ниже). Столбцы матрицы X соответствуют значениям факторов, а строки – веб-ресурсу. Выходной величиной y является порядок ранжирования результатов выдачи, т.е. номер сайта.

Для моделирования применяется обобщенный итерационный алгоритм ОИА МГУА [1], в котором матрица данных X делится на две части: первая (примерно 2/3 длины) – обучающая A, которая используется для оценки коэффициентов моделей, вторая (1/3 длины) – проверочная выборка B, на которой вычисляется качество модели как значение критерия регулярности AR:

$$AR = \left\| y_B - X_B \hat{\theta}_A \right\|^2, \qquad (1)$$

где $\hat{\theta}_A$ - оценка коэффициентов модели с помощью МНК.

Для моделирования процесса ранжирования результатов поиска веб-ресурсов были использованы следующие признаки (входные переменные):

x_1 – количество ключевых слов на сайте;

x_2 – количество ключевых слов на странице;

x_3 – отношение общего числа слов к числу ключевых слов на сайте;

x_4 – отношение всего числа слов к числу ключевых слов на странице;

x_5 – Google Page Rank (далее PR, результат расчета авторитетности веб-страниц по алгоритму системы);

x_6 – популярность тематики;

x_7 – число запросов по конкретному ключевому слову за определённый период времени;

x_8 – общее количество веб-страниц сайта;

x_9 – объём текста сайта;

x_{10} – объём сайта;

x_{11} – объём текста веб-страницы;

x_{12} – возраст сайта;

x_{13} – наличие ключевого слова в URL сайта (имя домена);

x_{14} – периодичность обновления информации на сайте;

x_{15} – последнее обновление страниц сайта;

x_{16} – число картинок (рисунков) на сайте;

x_{17} – количество мультимедийных файлов;

x_{18} – наличие замещающих надписей на рисунках (картинках);

x_{19} – длина (число символов) замещающих надписей рисунков (картинок);

x_{20} – использование фреймов;

x_{21} – язык сайта (русский или иностранный);
x_{22} – размер шрифта, которым оформлены ключевые слова;
x_{23} – жирность шрифта ключевых слов;
x_{24} – написаны ключевые слова в разрядку или нет;
x_{25} – написаны или нет ключевые слова заглавными буквами;
x_{26} – как далеко от начала веб-страницы располагаются ключевые слова;
x_{27} – наличие ключевых слов в заголовке;
x_{28} – наличие ключевых слов в мета-тэгах;
x_{29} – наличие файла «robot.txt»;
x_{30} – географическое месторасположение сайта;
x_{31} – комментарии внутри программного кода сайта;
x_{32} – к какому типу страниц относится каждая страница сайта: html или asp;
x_{33} – наличие в составе сайта flash модулей;
x_{34} – наличие в составе сайта веб-страниц с незначительными отличиями друг от друга;
x_{35} – соответствие ключевых слов сайта тому разделу каталога поисковой машины, в котором зарегистрирован сайт;
x_{36} – наличие «шумовых слов» («стоп-слов»);
x_{37} – общее количество гиперссылок сайта;
x_{38} – количество внутренних гиперссылок сайта;
x_{39} – количество внешних гиперссылок сайта;
x_{40} – глубина сайта;
x_{41} – количество внешних ссылок, содержащих в названии ключевые слова;
x_{42} – индекс цитирования Яндекс (ТИЦ).

Выходной переменной y является позиция веб-ресурса среди результатов ранжирования поисковой выдачи системы.

Точность построенной модели будем рассчитывать по формуле коэффициента детерминации:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} 100\%, \qquad (2)$$

где $\bar{y}$ – среднее значение, $\hat{y}_i$ – выход модели.

С применением программной реализации ОИА МГУА была построена следующая модель, которая описывает результаты ранжирования веб-ресурсов в данной поисковой системе:

$$\begin{aligned} y = {} & 3{,}24 + 2{,}71x_3 + 0{,}12x_4 + 0{,}00003x_7 - 2{,}69x_{12} + 0{,}012x_{22} - 14{,}8x_{27} - \\ & - x_{28} - 27{,}29x_{35} + 4x_{40} - 0{,}006x_{41} - 7{,}89x_5x_6 + 0{,}06x_{14}x_{15}^2 + 0{,}002x_{37}x_{38}x_{39} \end{aligned}. \qquad (3)$$

Показатели качества модели: $AR(A)$ = 2,48; $AR(B)$ = 3,51, R^2 =92%. Полученный процент означает, что эта модель лишь в четырех случаях из 50-ти ука-

зала другой порядковый номер ресурса по сравнению с Google. При этом результат вычисления по модели (3) округлялся до целого числа, которое и было ранговой позицией ресурса, см. таблицу 1, где показаны результаты применения модели уже к 100 первым сайтам.

Таблица 1 – Результаты ранжирования веб-ресурсов с помощью МГУА

Место в google.com.ua	Значения по модели МГУА	Результаты округления
1	1,23	1
2	1,89	2
3	4,01	4
4	4,21	4
5	4,89	5
6	6,02	6
7	6,78	7
8	8,00	8
9	8,52	9
10	9,33	9
…	…	…
21	21,23	21
22	22,49	23
23	22,85	23
…	…	…
32	33,56	34
33	33,56	34
34	33,68	34
…	…	…
57	57,22	57
58	58,15	58
…	…	…
99	98,95	99
100	99,56	100

Проанализировав структуру построенной модели, можно сделать вывод, что на ранжирование веб-ресурсов в поисковой системе *google* наибольшее влияние оказывают следующие 16 факторов:

x_3 – отношение общего числа слов к числу ключевых слов на сайте;

x_4 – отношение общего числа слов к числу ключевых слов на странице;

x_5 – Google PR;

x_6 – популярность тематики;

x_7 – число запросов по конкретному ключевому слову за определённый период времени;

x_{12} – возраст сайта;

x_{14} – периодичность обновления информации на сайте;

x_{15} – последнее обновление страниц сайта;

x_{22} – размер шрифта, которым оформлены ключевые слова;

x_{27} – наличие ключевых слов в заголовке;

x_{28} – наличие ключевых слов в мета-тегах;

x_{35} – соответствие ключевых слов сайта тому разделу каталога поисковой машины, в котором зарегистрирован сайт;

x_{37} – общее количество гиперссылок сайта;

x_{38} – количество внутренних гиперссылок сайта;

x_{39} – количество внешних гиперссылок сайта;

x_{40} – глубина сайта;

x_{41}– количество внешних ссылок, содержащих в названии ключевые слова;

Проанализировав эти факторы, можно сказать, что на ранжирование веб-ресурсов в поисковой системе *google* влияют в основном внешние факторы (x_5, x_6, x_7, x_{12}, x_{35}, x_{39}, x_{41}), а не внутренние.

Проверим правильность работы построенной модели (3) на других поисковых запросах:

- «рецепт омлета»;
- «купить ноутбук Киев»;
- «экспертные системы».

Полученные результаты отражены в таблице 2.

Таблица 2 – Результаты ранжирования веб-ресурсов в google.com.ua

Место в google.com.ua	Значения по модели МГУА		
	«рецепт омлета» / округленный результат	«купить ноутбук Киев» / округленный результат	«экспертные системы» / округленный результат
1	0,83 / 1	1,02 / 1	0,78 / 1
2	1,91 / 2	2,11 / 2	2,02 / 2
3	3,09 / 3	3,56 / 4	3,01 / 3
…	…	…	…
15	14,89 / 15	15,08 / 15	14,98 / 15
16	16,02 / 16	16,06 / 16	14,99 / 15
17	16,78 / 17	17,21 / 17	14,99 / 15
…	…	…	…
21	19,52 / 20	21,11 / 21	21,03 / 21
22	21,33 / 21	22,13 / 22	21,89 / 22
23	23,01 / 23	24,05 / 24	22,99 / 23
…	…	…	…

Продолжение таблицы 2

37	37,91 / 38	36,99 / 37	36,89 / 37
38	37,95 / 38	38,00 / 38	38,01 / 38
39	38,23 / 38	38,78 / 39	39,05 / 39
…	…	…	…
62	63,06 / 63	62,12 / 62	62,13 / 62
63	63,56 / 64	63,42 / 64	62,58 / 63
64	64,18 / 64	64,01 / 64	64,02 / 64
…	…	…	…
77	77,02 / 77	76,01 / 76	78,00 / 78
78	78,11 / 78	77,72 / 78	78,32 / 78
…	…	…	…
89	88,95 / 89	89,23 / 89	89,00 / 89
…	…	…	…
100	99,86 / 100	100,56 / 101	100,01 / 100
$\boldsymbol{R^2}$	87%	95%	93%

Из таблицы 2 видно, что найденная с помощью ОИА МГУА модель с высокой точностью повторяет ранжирование поисковой системы Google для совершенно разных поисковых запросов и может быть применена для дальнейшего изучения данного способа ранжирования.

2. Построение модели ранжирования поисковой выдачи Яндекс

В данной задаче моделируем процесс ранжирования веб-ресурсов поисковой выдачи Yandex (yandex.ua) для поисковой фразы «теплообмен».

Для эксперимента было отобрано первых 50 сайтов поисковой выдачи по данному ключевому запросу. Матрица данных X содержит 42 переменных-фактора, которые численно характеризуют каждый сайт. Столбцы матрицы X соответствуют значениям факторов, а строки – веб-ресурсу. Как и в первой задаче, качество модели вычислялось как значение критерия регулярности AR (1) при такой же пропорции разделения данных на две части A и B. Выходной переменной y является позиция веб-ресурса в поисковой выдаче.

С применением ОИА МГУА была построена следующая модель, описывающая порядок ранжирования веб-ресурсов в поисковой системе:

$$y = 7{,}12 + 1{,}01x_3 + 0{,}12x_4 + 0{,}000001x_7 - 2{,}69x_{12} + 8{,}12x_{22} + 2{,}79x_{27} + \\ + 0{,}001x_{28} - 48{,}19x_{35} - 2{,}001x_{41} - 12{,}22x_{42}x_6 - 3{,}08x_{14}x_{15}^2 + 0{,}04x_{37}x_{38}x_9 \quad (4)$$

Показатели качества модели: $AR(A) = 3{,}12$; $AR(B) = 3{,}92$, $R^2 = 89\%$.

Таблица 3 – Результаты ранжирования веб-ресурсов в yandex.ua

Место в yandex.ua	Значения по МГУА	Результаты округления
1	0,83	1
2	1,29	2
3	3,08	3
4	5,01	5
5	5,23	5
6	6,02	6
7	7,78	8
8	8,09	8
9	9,22	9
10	10,13	10
…	…	…
21	21,23	21
22	21,99	22
23	23,85	24
…	…	…
32	32,56	33
33	33,07	33
34	34,12	34
…	…	…
57	57,02	57
58	58,11	58
…	…	…
99	99,95	100
100	107,12	107

Из структуры модели следует, что на ранжирование веб-ресурсов в поисковой системе Yandex наибольшее влияние оказывают следующие 13 факторов:

x_3 – отношение общего числа слов к числу ключевых слов на сайте;

x_4 – отношение общего числа слов к числу ключевых слов на странице;

x_6 – популярность тематики;

x_7 – число запросов по ключевому слову за определённый период времени;

x_{12} – возраст сайта;

x_{14} – периодичность обновления информации на сайте;

x_{15} – последнее обновление страниц сайта;

x_{22} – размер шрифта, которым оформлены ключевые слова;

x_{27} – наличие ключевых слов в заголовке;

x_{28} – наличие ключевых слов в мета-тэгах;

x_{35} – соответствие ключевых слов сайта тому разделу каталога поисковой машины, в котором зарегистрирован сайт;

x_{41}– количество внешних ссылок, содержащих в названии ключевое слова;

x_{42} – индекс цитирования Яндекс.

Проанализировав эти факторы, можно сказать, что на ранжирование веб-ресурсов в поисковой системе yandex влияют в основном внешние факторы (x_6, x_7, x_{12}, x_{35}, x_{41}, x_{42}).

Проверим правильность работы найденной формулы (4) на других поисковых запросах:

- «теория вероятности»;
- «химчистка ковров»;
- «отдых в Таиланде».

Таблица 4 – Результаты ранжирования веб-ресурсов в yandex.ua

Место в yandex.ua	Значения по МГУА		
	«теория вероятности» / округленный результат	«химчистка ковров» / округленный результат	«отдых в Таиланде» / округленный результат
1	0,95 / 1	1,12 / 1	1,18 / 1
2	1,91 / 2	2,11 / 2	2,00 / 2
3	3,21 / 3	3,46 / 4	3,61 / 4
…	…	…	…
15	15,09 / 15	15,08 / 15	14,18 / 14
16	16,00 / 16	16,08 / 16	14,99 / 15
17	17,12 / 17	17,21 / 17	15,09 / 15
…	…	…	…
21	20,52 / 21	21,11 / 21	21,03 / 21
22	22,33 / 22	22,13 / 22	22,79 / 23
23	23,01 / 23	23,95 / 24	23,99 / 24
…	…	…	…
37	37,51 / 38	36,99 / 37	37,12 / 37
38	37,95 / 38	38,00 / 38	38,01 / 38
39	39,23 / 39	38,78 / 39	38,05 / 38
…	…	…	…
62	63,06 / 63	62,12 / 62	61,93 / 62
63	63,56 / 64	63,42 / 64	62,58 / 63
64	64,18 / 64	64,01 / 64	64,47 / 65
…	…	…	…
77	77,02 / 77	77,01 / 77	78,00 / 78

Продолжение таблицы 4

78	78,01 / 78	78,22 / 78	78,82 / 79
…	…	…	…
89	88,95 / 89	89,23 / 89	88,00 / 88
…	…	…	…
100	99,86 / 100	100,06 / 100	99,01 / 99
R^2	85%	88%	84%

Из таблицы 4 видно, что построенная модель с высокой точностью соответствует результатам ранжирования поисковой системы Yandex.

Сравнив полученные модели ранжирования для Google (3) и Яндекс (4), видим, что из-за различных алгоритмов расчета авторитетности веб-страницы (PR для Google и ТИЦ для Яндекс) они отличаются только такими признаками:

x_{40} – возраст домена;

x_5 – значение PR;

x_{42} – значение ТИЦ.

Остальные признаки входят в обе модели, но с разными коэффициентами.

3. Выводы

Анализ полученных в ходе исследования моделей ранжирования поисковых систем Google и Яндекс показал следующее:

- в обе модели входят почти одни и те же признаки ранжирования, и разница в них заключается в основном в коэффициентах при этих признаках;
- в обеих моделях преимущественно используются внешние (более помехоустойчивые) признаки ранжирования, которые сложнее искусственно накручивать при продвижении произвольного сайта;
- основные отличия моделей ранжирования заключаются в алгоритмах расчета авторитетности веб-страниц, которые являются запатентованными разработками и держатся в строгом секрете [2].

Исследование показало, что, благодаря независимости признаков ранжирования от смыслового наполнения сайта, в рамках одной поисковой системы можно использовать одну и ту же модель ранжирования для запросов из совсем разных областей знаний.

Высокая точность построенных моделей ранжирования доказывает эффективность применения ОИА МГУА для решения подобного рода задач.

Література

1. Степашко В.С., Булгакова О.С., Зосімов В.В. Гібридні алгоритми самоорганізації моделей для прогнозування складних процесів. – Індуктивне моделювання складних систем. Зб. наук. праць, вип. 2. – К.: МННЦ ІТС НАН та МОН України, 2010.– С. 236-246.
2. Колисниченко Д.Н. Поисковые системы и продвижение сайтов в Интернете. — М.: «Диалектика», 2007. — 272 с.