5. Web-источник: Терминологический словарь обучения – Проектирование автоматизации учебных курсов [18.01.2011] http://www.dupliksv.hut.ru/pauk/dict/
6. *Гладышев П.Е., Сиговцев Г.С.* Модель адаптивного учебного интернет-ресурса. // XI Всероссийская научно-методическая конференция "Телематика'2004".

Поступила 24.03.2011р.

УДК 004.832.3

В.В. Зосимов, аспирант, отдел распределенных интеллектуальных систем МНУЦИТиС НАН и МОН Украины
А.С. Булгакова, аспирант, отдел индуктивного моделирования МНУЦИТиС НАН и МОН Украины

МОДЕЛИРОВАНИЕ ПРОЦЕССА РАНЖИРОВАНИЯ РЕЗУЛЬТАТОВ ПОИСКОВОЙ ВЫДАЧИ ВЕБ-СТРАНИЦ В СЕТИ ИНТЕРНЕТ

*Аннотация***.** В данной статье описано моделирование процесса ранжирования результатов поисковой выдачи Google при помощи GIA GMDH (Обобщенный итерационный алгоритм МГУА). В ходе процесса моделирования были выбраны наиболее значимые для построения поисковой выдачи параметры веб-страниц. В результате исследования мы получили математическое отображение (модель) процесса ранжирования результатов поисковой выдачи веб-страниц в сети Интернет.

Abstract. This article describes modeling of ranking search engine results in Google using the GIA GMDH (generalized iterative algorithm for GMDH). During the simulation there were chosen the most important search parameters of web pages for building of related search results. As a result of our research we have got a mathematical mapping (model) of the process of ranking the results of search results of web pages on the Internet.

*Ключевые слова***.** Алгоритм, ранжирование сайтов, МГУА, обобщенный итерационный алгоритм, моделирование.

Введение. Мы уже давно привыкли находить почти любую, интересующую нас информацию в Интернете. Это стало возможным во многом благодаря появлению и развитию поисковых систем, основная задача которых – сделать поиск информации в Интернете более эффективным. Для этого были разработаны алгоритмы ранжирования сайтов в поисковой выдаче, результатом работы которых является отсортированный по релевантности (наиболее точно удовлетворяющий условия поиска) список сайтов.

Современный Интернет все больше становится похож на огромную рекламную площадку, что значительно затрудняет поиск информации в сети.

Алгоритмы работы поисковых систем развиваются вместе с Интернетом, они постоянно совершенствуются, но, не смотря на это, поиск нужной информации становится все более сложной задачей.

В ходе исследования было проведено моделирование процесса ранжирования по нескольким поисковым запросам. В данной работе описан результат моделирования процесса ранжирования поисковой выдачи веб-страниц в сети Интернет по запросу «Защита информации».

Объектом исследования являются процессы поиска информации в сети Интернет.

Предмет исследования – методы работы поисковых систем и алгоритмы ранжирования сайтов в поисковой выдаче.

Цель исследования – получить математическую модель процесса ранжирования поисковой выдачи веб-страниц в сети Интернет.

Ожидаемые результаты – эффективная математическая модель, позволяющая провести более полный анализ процесса ранжирования поисковой выдачи веб-страниц в сети Интернет.

Основное направление нашей деятельности – анализ существующих алгоритмов ранжирования сайтов в поисковой выдаче с целью выявления их недостатков и дальнейшей их оптимизации.

1. Набор факторов, использованных для моделирования процесса ранжирования результатов поисковой выдачи веб-страниц в сети Интернет

Ранжирование документов в поисковых машинах - процесс весьма и весьма сложный. Разработчики постоянно пытаются совершенствовать алгоритмы ранжирования, преследуя, как правило, две большие цели - улучшение качества поиска и уменьшение возможности искусственных воздействий на ранжирование результатов. Та или иная поисковая машина может учитывать множество факторов, так или иначе влияющих на положение конкретного документа в выдаче по конкретному запросу. Большую часть своих достижений в области ранжирования документов разработчики поисковых алгоритмов хранят в строгом секрете, ограничиваясь публикациями либо каких-то весьма общих фактов, либо, наоборот, описанием очень частных задач, возможно, чрезвычайно интересных с точки зрения разработчика, но мало полезных на практике тем, что пытается улучшить ранжирование конкретного сайта по конкретным запросам. Специалисты в области SEO, поэтому, очень ограничены в информации и могут добывать ее только экспериментальным путем, оценивая работу поисковых алгоритмов путем построения так называемой модели "чёрного ящика" с известными выходными и входными параметрами и неизвестным внутренним устройством. Манипулируя входной информацией, т.е. изменяя для конкретных документов факторы, которые учитываются при ранжировании, и оценивая изменение выходной информацией, т.е. положением этих документов в выдаче по конкретным

запросам, можно сделать определенные выводы о том, какие факторы и каким образом учитываются поисковыми машинами. Это знание позволит сформировать оптимальную стратегию продвижения ресурса в поисковых машинах с целью привлечения максимального количества целевых посетителей при минимальных затратах.

Механизм ранжирования - это программа, которая определяет релевантность страницы (степень соответствия) поисковому запросу на основе семантического анализа документа, плотности и соответствия ключевых слов, ссылок с других Интернет-ресурсов и других параметров. От релевантности страницы зависит ее место при выводе результатов поиска.

Факторы, которые используют современные поисковые системы при ранжировании поисковой выдачи, делятся на внешние и внутренние.

Внешние факторы направлены на увеличение авторитетности сайта, его «общего веса». Их совокупность не влияет на релевантность сайта напрямую, но является важным коэффициентом при формировании поисковой выдачи.

К внутренним факторам относятся все внутреннее наполнение сайта. Поисковые системы в ходе индексации страниц сайта проводят семантический анализ его содержимого и на основе полученных данных определяют релевантность сайта тем или иным поисковым запросам.

2. Моделирование процесса ранжирования

2.1. Описание алгоритма. Для моделирования процесса ранжирования веб-ресурсов был использован обобщенный алгоритм МГУА (рис 1). Предложенный алгоритм дает возможность не только усовершенствовать генератор структур классического многорядного алгоритма МГУА и получить новые варианты, которые позволяют не терять информативные аргументы, что могут быть отсеяны на предыдущих этапах моделирования.

Описание алгоритма:

Первая итерация (первый ряд):

Шаг 1. Из множества входов $X = \{x_1, x_2, ..., x_m\}$ выбираются пары аргументов x_i, x_j и формируются описания частных моделей вида

$$Y_l^{(1)} = \varphi(x_i, x_j),\ i \neq j,\ ((i, j = 1,2,...,m),$$

при этом используют следующие частные описания:

$$Y_l^{(1)} = a_0 + a_1 x_i + a_2 x_j$$

или

$$Y_l^{(1)} = a_0 + a_i x_i + a_j x_j + a_{ij} x_i x_j + a_{ii} x_i^2 + a_{jj} x_j^2 \qquad (1)$$

Шаг 2. Для каждой частной модели используется комбинаторная оптимизация, то есть для каждого сгенерированной пары, применяется комбинаторный алгоритм МГУА для выбора лучшей модели. Общая модель, которая используется для комбинаторного перебора имеет вид

$$f(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i^2 + a_4 x_j^2 + a_5 x_i x_j . \qquad (2)$$

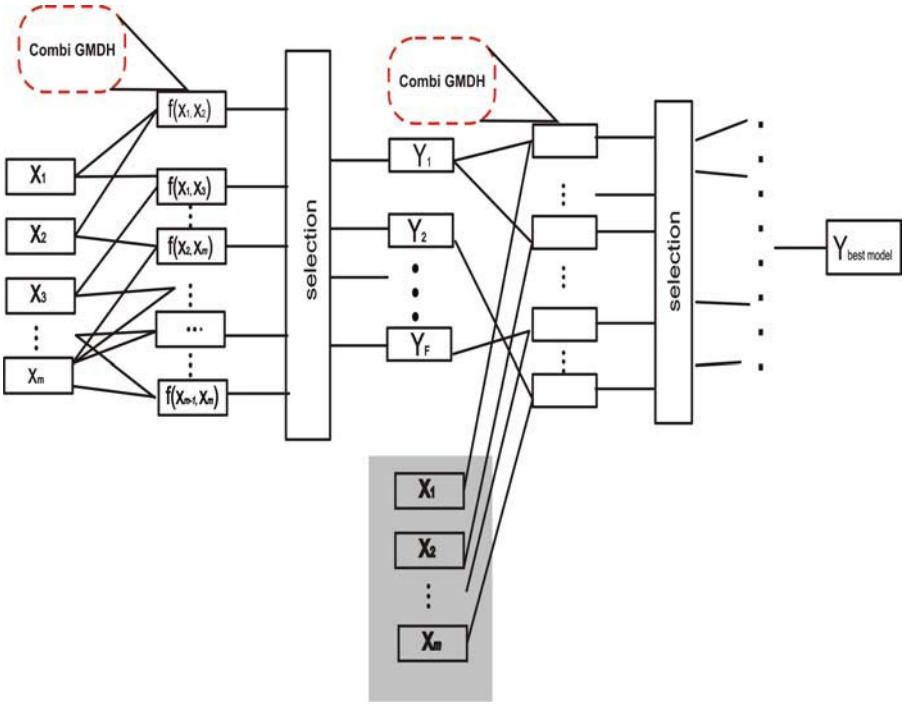

Рис. 1. Обобщенный алгоритм МГУА

*Шаг 3***.** Используя метод наименьших квадратов (МНК) для каждого описания находятся по учебной выборке оценки неизвестных коэффициентов

$$\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5, .$$

Шаг 4. По минимальным значением критерия на проверочной последовательности отбираем *F* лучших моделей, то есть реализуя процедуру селекции. Выходы этих моделей служат аргументами-входами для конструирования моделей второго ряда.

Шаг 5. Находится значение критерия ряда

$$C(0) = \min_l C_l$$

r-ая итерация (ряд r):

Шаг 1. Формируются описания частных моделей вида 1.

Шаг 2.

Для каждой частной модели используется комбинаторная оптимизация, то есть для каждого сгенерированной пары, применяется комбинаторный алгоритм МГУА для выбора лучшей модели. Общая модель, которая используется для комбинаторного перебора имеет вид

$$F(y_l^r, y_{l-1}^r) = a_0 + a_1 y_i^{r-1} + a_2 y_j^{r-1} + a_3 \left(y_i^{r-1}\right)^2 + a_4 y_i^{r-1} y_j^{r-1} + a_5 \left(y_j^{r-1}\right)^2 . \quad (3)$$

Шаг 3. Используя метод наименьших квадратов (МНК) для каждого

описания находятся по учебной выборке оценки неизвестных коэффициентов $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5,$.

Шаг 4. По проверочной выборке находится для каждого частного описания величина критерия C^r.

Шаг 5. Находится $C^r = \min_l C_l^r$. Проверяется условие $C^r > C^{r-1}$, где C^r, C^{r-1} – величины критерия точности для наилучших моделей (*r-1*)-го и *r-го* ряда селекции соответственно. Если так, то конец. Искомая модель выбирается из частных описаний *r-1-го* уровня, на котором достигается минимальное значение критерия C^{r-1}. Иначе переход к следующему ряду.

Заключительный этап:

Двигаясь от конца к началу и делая последовательную замену переменных, вычисляются выражения для искомой модели в начальном пространстве описаний.

2.2. Описание и результаты эксперимента. Выборка содержит 31 переменную и делится на две части: 2/3 – учебная А, которая используется для оценки коэффициентов, другая 1/3 – тестовая выборка В.

Качество модели было вычислено выборке В, как значение критерия регулярности, AR.

$$AR = \left\| y_B - X_B \hat{\theta}_A \right\|^2 \quad (4)$$

Входные переменные:

Для моделирования процесса ранжирования результатов поиска веб-страниц мы будем использовать следующие факторы:

- x1 – наличие искомой фразы в тэге TITLE.
- x2 - наличие искомой фразы в тэгах H1-H6.
- x3 - наличие искомой фразы в тэге STRONG.
- x4 - наличие искомой фразы в тэге DESCRIPTION.
- x5 - наличие искомой фразы в тэге KEYWORDS.
- x6 – значение PR главной страницы сайта.
- x7 – наличие прямого вхождения искомой фразы в тексте страницы.
- x8 – наличие словоформ искомой фразы с сохранением части речи.
- x9 – наличие словоформ искомой фразы с изменением части речи.
- x10 – соответствие регистра искомой фразы.
- x11 – % уникального контента.
- x12 – близость искомой фразы к началу страницы.
- x13 – плотность вхождения искомой фразы в текст страницы (измеряется в процентах).
- x14 – ошибки кодировки.
- x15 – отсутствие на сайте значительного количества 404 ошибок.

- x16 – высокий аптайм сервера.
- x17 – высокая скорость загрузки страницы.
- x18 – наличие искомой фразы в имени домена.
- x19 – количество внешних ссылок на страницу.
- x20 – дата последней индексации.
- x21 – возраст домена (количество дней, прошедших со дня его регистрации).
- x22 – возраст URL страницы.
- x23 – динамика появления контента на сайте.
- x24 – наличие искомой фразы в тэге ALT картинок, расположенных на странице.
- x25 – наличие искомой фразы в названии картинок, расположенных на странице.
- x26 – наличие искомой фразы в тэге TITLE картинок, расположенных на странице.
- x27 – количество исходящих ссылок.
- x28 – наличие карты сайта.
- x29 – указание приоритета индексации страниц в карте сайта.
- x30 – небольшое количество внешних ссылок со страницы донора.
- x31 – плавность динамики прироста внешних ссылок.
- x32 – посещаемость сайта из других источников.
- x33 – наличие счетчика.

Выходная переменная: y – позиция веб-ресурса среди результатов ранжирования поисковой выдачи.

Используя программу, которая реализует работу обобщенного алгоритма МГУА [2], была получена следующая формула, которая описывает процесс ранжирования веб-ресурсов в поисковой системе:

$$y = 3{,}000 + -0{,}823x_6 + 0{,}0001x_{13} - 0{,}0001x_{21} - 0{,}035x_{24} - 0{,}023x_6x_{21}$$
$$+1{,}2643x_6x_1 + 0{,}0012x_{19}x_{30}^2 + 0{,}011x_{29}x_{28}^2$$

AR(B) = 2,48

Проанализировав полученную модель можно сделать вывод, что на ранжирование веб-ресурсов в поисковой системе влияют следующие факторы:

- значение PR главной страницы сайта.
- плотность ключевых слов на странице.
- возраст домена.
- наличие искомой фразы в тэге ALT картинки.
- наличие карты сайта.
- указание приоритета индексации страниц в карте сайта.
- небольшое количество внешних ссылок со страницы донора.
- количество внешних ссылок на страницу.

Заключение. На основе полученных в результате исследования данных планируется разработать методы, которые позволят оптимизировать существующие алгоритмы ранжирования сайтов в поисковой выдаче. Это позволит сделать поиск информации в сети Интернет более простым и удобным, что значительно сэкономит время пользователей сети.

1. *Volodymyr Stepashko, Oleksandra Bulgakova, Viacheslav Zosimov*. Modified multilayered GMDH algorithm with combinatorial optimization of partial descriptions complexity. – Proceedings of the International Workshop on Inductive Modelling IWIM-2010, Ukraine. – Yevpatoria, 2010.
2. *Зосимов В.В., Булгакова А.С.* Проектирование программного модуля для исследования работы алгоритмов поисковых систем.
3. Байков В.Д. Интернет. Поиск информации. Продвижение сайтов. — СПб.: БХВ-Петербург, 2000. — С. 288.
4. *Колисниченко Денис Николаевич* Поисковые системы и продвижение сайтов в Интернете. — М.: «Диалектика», 2007. — С. 272.
5. *Ашманов И.С., Иванов А.А.* Продвижение сайта в поисковых системах. — М.: «Вильямс», 2007. — С. 304.
6. *Ivakhnenko A.G*. Group method of data handling - competitor for the method of stochastic approximation, Soviet Automatic Control, No. 3, pp. 58-72, 1968.
7. *Stepashko V.S.* Combinatorial GMDH algorithm with the optimal scheme of models sorting-out, Soviet Automatic Control, No. 3, pp. 31-36, 1981.
8. *Bulgakova O., Kordik P.* Methods of true data mining model selection - with experimental results. Proceedings of IWIM 2009 in Krynica, Poland, pp. 23-27, 2009.

Поступила 17.03.2011р.

УДК 004.3

Н.Є. Фролова, Національний авіаційний університет, Київ
О.О. Бакіна, Національний авіаційний університет, Київ

ПЕРСПЕКТИВИ МІНІМІЗАЦІЇ ЛАТЕНТНОСТІ ОПЕРАТИВНОЇ ПАМ'ЯТІ

Basic progress of technologies of main memory trends are analyzed. The features of the modules of memory of DDR3 are considered in relation to the increase of carrying capacity due to minimization of delays.

Вступ. Зі збільшенням пропускної здатності оперативної пам'яті значно збільшуються затримки доступу до пам'яті, що призводить до простою процесора. Тому пошук альтернативних шляхів прискорення роботи з пам'яттю, а саме знаходження та реалізація способів зменшення латентності є найактуальнішою проблемою, що стоїть перед розробниками.

