

УДК 004.89

РЕГРЕССИОННЫЙ АНАЛИЗ АЛГОРИТМА РАНЖИРОВАНИЯ РЕЗУЛЬТАТОВ ТЕКСТОВОГО ПОИСКА В БАЗАХ ДАННЫХ СИСТЕМ ГЕОМОНИТОРИНГА С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Тарков Михаил Сергеевич,

канд. техн. наук, ст. науч. сотр. лаборатории физических основ
материаловедения кремния Института физики полупроводников
им. А.В. Ржанова СО РАН, Россия, 630090, Новосибирск,
пр. Ак. Лаврентьева, 13. E-mail: tarkov@isp.nsc.ru

Кожушко Оюна Алексеевна,

аспирант Новосибирского государственного университета,
Россия, 630090, г. Новосибирск, ул. Пирогова, 2. E-mail: oyuna@mail.ru

Актуальность исследования обусловлена необходимостью изучения поведения алгоритмов текстового ранжирования. Практическую ценность исследование представляет для разработчиков поисковых систем, в том числе при решении задач распознавания и адаптивной классификации объектов по данным спутниковых систем геомониторинга.

Цель исследования: изучение нейросетевой модели алгоритма ранжирования текстовых документов в базах данных систем геомониторинга, построенной на основе использования сети Кохонена, многослойных персептронов и метода кластеризации *k-means*.

Методы исследования: программная реализация и тестирование нейросетевых алгоритмов ранжирования путем сравнения результатов их работы с результатами классического алгоритма ранжирования *OkapiBm25*.

Результаты исследования. Предложен алгоритм идентификации систем текстового поиска в базах данных систем геомониторинга, построенный на основе нейросетевых моделей обработки данных и включающий факторный и регрессионный анализ данных. Факторный анализ включает кластеризацию данных на основе использования сети Кохонена. Для упрощения обучения разработан алгоритм факторного анализа, исключающий малозначимые для ранжирования характеристики. Результатом работы моделей является набор значимых при ранжировании характеристик и их оптимальные значения. Для проведения регрессионного анализа предлагается использовать одну из двух нейросетевых моделей: на основе гибридной нейронной сети или на основе комплекса многослойных персептронов. Выбор модели регрессионного анализа осуществляется на основе результатов кластерного и факторного анализа. В случае выделения большого числа кластеров входных векторов предпочтительнее использование модели на основе гибридной нейронной сети. В случае слабых пересечений наборов значимых характеристик между кластерами предпочтительнее использование модели на основе комплекса многослойных персептронов. Результаты тестирования алгоритма показывают успешное обучение моделей и низкие значения ошибок обучения и тестирования. Предложенные модели апробированы на тестовых данных алгоритма семейства *OkapiBm25*, и выявлены особенности их применения в зависимости от характеристик входных данных.

Ключевые слова:

Базы данных систем геомониторинга, алгоритм текстового ранжирования, регрессионный анализ, факторный анализ, классификация, кластеризация, нейронные сети, сеть Кохонена, многослойный персептрон.

Введение

Эффективный поиск документов в базах данных систем геомониторинга [1] с каждым годом превращается во все более сложную задачу. Причиной этого являются быстрый рост объема информации и появление новых особенностей коллекции веб-документов. Современные поисковые системы постоянно оптимизируют свою работу, уделяя большое внимание ранжированию найденных документов по степени их релевантности запросу [2]. Для решения данной задачи в настоящее время активно используется машинное обучение системы поиска по ассессорским базам знаний [3] или на основе статистической информации [4, 5]. Такой подход позволяет достичь высокой эффективности, однако делает алгоритм ранжирования «черным ящиком», что приводит к проблеме идентификации системы и поиска ее ключевых элементов [6]. Многие исследователи сконцентрированы на проблеме повышения качества поиска, мало затрагивая задачу идентификации алгоритма ранжиро-

вания. Идентификация алгоритма позволяет: оценить поведение алгоритма ранжирования в зависимости от разных входных данных, выявить аномалии в работе алгоритма и возможности повышения качества за счет оптимизации использования отдельных факторов. Кроме сферы информационного поиска результаты решения данной задачи востребованы в области интернет-маркетинга в направлении поисковой оптимизации веб-сайтов [7]. Изучение систем ранжирования также важно при дистанционном зондировании Земли [8].

Ранее для анализа алгоритма ранжирования применялись статистические методы [9], показывавшие удовлетворительные результаты для явных алгоритмов, таких как алгоритм текстового ранжирования Яндекс *Atr* [10]. С началом использования скрытых алгоритмов появилась необходимость изменения методов анализа и применение методов машинного обучения. И. Зябровым, О. Пожарковым и И. Пожарковой предложены: алгоритм оценки оптимальных значений некоторых

характеристик документов с помощью генетических алгоритмов [11, 12], метод моделирования алгоритма ранжирования Яндекс с помощью технологии MatrixNet [13]. Эти методы позволили выявить важные особенности алгоритма, но не обладают возможностью дообучения, что ограничивает их применимость в условиях частых изменений настроек алгоритма и расширения обучающих данных. В данной статье задача анализа алгоритма ранжирования решается с помощью нейросетевого подхода, позволяющего производить дообучение системы [14–16]. Ранее для решения этой задачи нейронные сети не применялись, однако известны примеры их успешного применения для других задач идентификации [17] и анализа текстовых документов [18].

Постановка задачи

Анализ скрытого алгоритма ранжирования является задачей интеллектуального анализа данных (Data Mining) [19], для решения которой необходимо выявить факторы, оказывающие существенное влияние на результат ранжирования, и зависимость позиций документов от значений этих факторов. Исследуемые данные представляются в векторном виде, где компонентам векторов соответствуют факторы, характеризующие документы и запросы. Примерами таких факторов являются длина документа, цитируемость документа, количество слов в запросе и другие.

Алгоритм ранжирования поисковой системы строит функцию релевантности, которая сопоставляет паре векторов (q, d) , описывающих документ и текстовый запрос соответственно, числовую оценку релевантности rel [20]:

$$f(q, d) \rightarrow rel.$$

Решение задачи идентификации алгоритма ранжирования сводится к построению модели, которая по заданным векторам (q, d) определяет степень релевантности документа d запросу q . Степень релевантности определяется как ранг, присваиваемый документу d при ранжировании по запросу q . Обычно разделяют несколько степеней релевантности (например, высокорелевантный, среднерелевантный и низкорелевантный) и каждой степени сопоставляется несколько рангов.

В данной работе рассмотрен частный случай описанной задачи идентификации. Требуется определить значимые факторы и их значения, достаточные для определения принадлежности или непринадлежности документа классу документов, которые релевантны заданному запросу. Таким образом, в задаче рассматривается только одна степень релевантности, и решение задачи позволит ответить на вопрос, какими свойствами должен обладать документ, релевантный заданному запросу. Исходя из такой постановки задачи идентификации, на вход системе подается запрос, а на выходе система предоставляет значения значимых факторов.

Подготовка данных

Факторы, описывающие документ и текстовый запрос, могут быть как числовыми (такие, как объем текста документа), так и номинальными (например, тематика документа). Целью предварительной обработки данных для нейросетевого анализа является приведение данных к однородному виду, которое включает в себя три этапа [14, 16]:

1. Априорное исключение малозначимых компонент векторов d и q . На данном этапе проверяется наличие корреляционных связей между характеристиками как входных, так и выходных данных отдельно с последующим исключением малозначимых факторов.
2. Представление входов и выходов в числовом виде для номинальных факторов с помощью двоичного кодирования, при котором каждому значению фактора сопоставляется вектор, компоненты которого соответствуют разрядам двоичного представления номера значения.
3. Нормировка данных с помощью биполярной сигмоидальной функции активации [16]

$$f(x) = \tanh(\beta x),$$

где β – заданный коэффициент; x – значение фактора.

Нейросетевые модели

Поисковая система анализирует все известные ей факторы, однако в реальном расчете релевантности используются только наиболее значимые из них [3]. Это делает необходимым проведение факторного анализа данных для выявления значимых факторов. Часто для различных пар (q, d) алгоритм ранжирования использует различные факторы. В данной работе значимыми считаются факторы, принимающие близкие выходные значения для схожих входных векторов. В качестве входных векторов рассматриваются векторы x запросов q , в качестве выходных векторов y – векторы d характеристик документа. Также может рассматриваться модель, в которой часть компонент вектора d входит во входной вектор x , а оставшаяся часть составляет выходной вектор y .

Предложенный ниже алгоритм факторного анализа использует сеть Кохонена [14–16] и метод кластеризации k-means [14, 16]. Использование сети Кохонена обуславливается сокращением временных издержек на переобучение и возможностью использования результатов обучения на этапе регрессионного анализа.

Метод кластеризации k-means состоит в следующем. Пусть задано множество векторов $\{x_1, \dots, x_m\} \subseteq E$, E – пространство векторов. Требуется разбить это множество на k кластеров. Для этого находится множество кластерных центров $w_j \subseteq E$, $j=1, \dots, k$, минимизирующих функционал

$$D = \sum_{i,j} d(x_i, w_j),$$

где $d(x_i, w_j)$ – мера расстояния между векторами x_i и w_j в пространстве E .

Нейронная сеть Кохонена состоит из одного слоя – выходного (рис. 1).

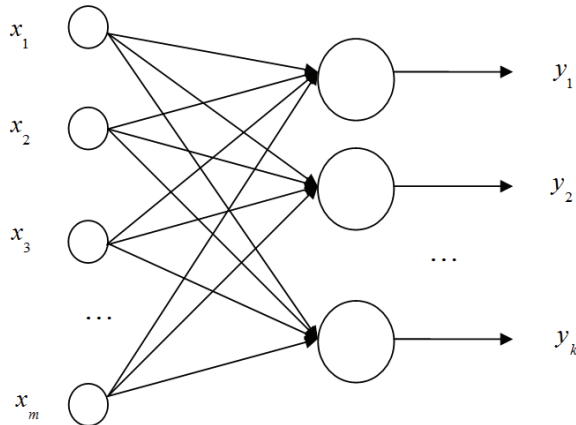


Рис. 1. Архитектура нейронной сети Кохонена

Fig. 1. Architecture of Kohonen neural network

Вектор входных сигналов $x=(x_1, \dots, x_m)$ поступает на входы всех нейронов выходного слоя. Нейроны выходного слоя являются линейными адаптивными сумматорами и обучаются по правилу WTA (Winner Takes All). Каждому нейрону выходного слоя соответствует некоторый кластер пространства E .

Обучение сети Кохонена состоит в подборе весов w_{ij} , $i=1 \dots m$, $j=1 \dots k$ нейронов выходного слоя. Веса нейрона, отвечающего за кластер, являются компонентами вектора w^j , являющегося центром j -го кластера. Нейрон $j \in \{1, \dots, k\}$, победивший при обучении на некотором входном векторе, изменяет свои веса согласно правилу

$$w^j(t+1) = w^j(t) + \eta(x - w^j(t)),$$

где w^j – вектор весовых коэффициентов нейрона-победителя; $\eta \in (0, 1)$ – коэффициент обучения; x – входной вектор. Одновременно с нейроном-победителем с меньшей интенсивностью свои веса меняют и другие нейроны $i \in \{1, \dots, k\}$, $i \neq j$. Чем дальше нейрон находится от победителя, тем меньше изменяются его веса:

$$w^i(t+1) = w^i(t) + \eta G(i, j)(x - w^i(t)).$$

Здесь $G(i, j)$ – функция расстояния между нейроном-победителем j и данным нейроном i , например

$$G(i, j) = \exp\left(-\frac{d^2(i, j)}{2\lambda^2}\right),$$

где $d(i, j)$ – мера расстояния между весовыми векторами нейронов (например, $d(i, j) = \|w^i - w^j\|$), λ – заданный параметр. Выходной сигнал нейрона i равен $y_i = G(i, j)$, $i=1, \dots, k$.

Алгоритм факторного анализа. На вход алгоритма поступает множество входных векторов q и множество выходных векторов d . Алгоритм факторного анализа состоит из двух шагов:

1. Выполнить многомерную кластеризацию входных векторов x с помощью сети Кохонена. Ко-

личество k выходных нейронов (количество кластеров) определить экспериментально, проводя обучение необходимое число раз.

2. Для каждого кластера и каждой компоненты y_j , $j=1, \dots, k$ выходного вектора сети Кохонена провести одномерное разбиение значений x на два кластера методом k -means. Исходя из результата разбиения и заданных параметров ε и p , принять решение о значимости характеристики y_j по следующему правилу: если доля векторов, находящихся в меньшем кластере, превышает $p \in (0, 1)$, и

$$\sum_{i,l} d(x^i, w^l) > \varepsilon,$$

где w^l , $l=1, 2$ – центры двух кластеров; x^i – значения фактора; $i=1, \dots, N$, N – количество примеров; $d(x, w)$ – мера близости векторов x и w , признать характеристику y_j незначимой. Использование параметров ε и p позволяет отличать наличие одного кластера от наличия двух близко расположенных кластеров.

По завершении факторного анализа получены:

- 1) разбиение входных векторов x на k кластеров и центры w^j , $j=1 \dots k$ этих кластеров;
- 2) обученная сеть Кохонена;
- 3) список значимых характеристик (факторов) для каждого кластера.

Регрессионный анализ. В ходе регрессионного анализа требуется получить оптимальные значения значимых характеристик. В качестве модели регрессионного анализа рассмотрены две конструкции, в основе которых лежит многослойный персептрон [14–16].

Многослойный персептрон. Каждый нейрон персептрона получает на вход вектор $x=(x_1, \dots, x_N)$ и

генерирует выходной сигнал $y = f\left(\sum_{i=0}^N w_i x_i\right)$, где w_i , $i=1 \dots N$ – настраиваемые веса; w_0 – пороговое значение, а $x_0=1$. В случае, когда функция $f(u) = \frac{1}{1 + e^{-\beta u}}$

(униполярная сигмоидальная функция) или $f(u) = \tanh(\beta u)$ (биполярная сигмоидальная функция), где β – заданный коэффициент, многослойная нейронная сеть называется сигмоидальной. Такая модель используется в данной работе.

Множество нейронов многослойной нейронной сети разделено на выходной и заданное количество скрытых слоев. Количество входов персептрона, образующих входной слой, и нейронов выходного слоя полностью определяется условием решаемой задачи, в то время как количество нейронов скрытых слоев определяется экспериментально. Нейроны каждого слоя обычно связаны со всеми нейронами соседних слоев. Например, в сигмоидальной сети, представленной на рис. 2, нейроны скрытого слоя связаны со всеми входами сети (входной слой) и всеми нейронами выходного слоя. Сеть является однонаправленной, то есть сигналы передаются от входов сети в направлении к выходным нейронам.

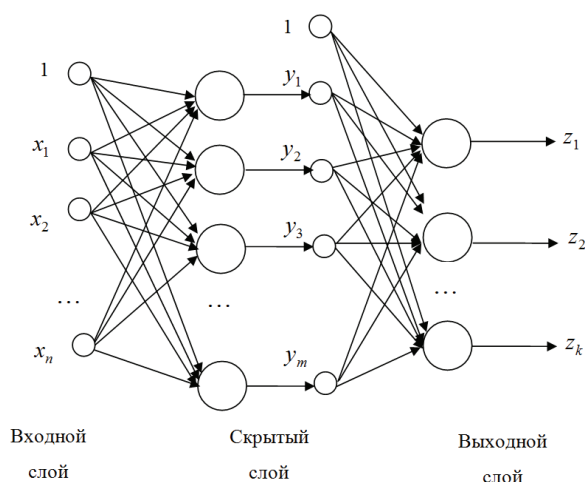


Рис. 2. Сигмоидальная сеть с одним скрытым слоем

Fig. 2. Sigmoid network with one hidden layer

Обучение нейрона сети состоит в подборе весов связей – коэффициентов w_i , $i=1...N$. Обучение нейронной сети является задачей оптимизации целевой функции ошибки при варьировании весовых коэффициентов. Функция ошибки может быть задана различными способами, например, среднеквадратичная функция ошибки

$$E(w) = \frac{1}{2} \sum_i \sum_j (z_{ij} - d_{ij})^2,$$

где z_{ij} – компонента выходного вектора, продуцируемая j -м выходным нейроном сети в ответ на i -й входной вектор задачника; d_{ij} – компоненты целевого вектора для i -го входного вектора задачника.

Поскольку сигмоидальная функция обладает свойством дифференцируемости, возможно применение градиентных методов обучения. В данной работе используется метод сопряженных градиентов [16]. Начальное значение весов нейронов выбирается случайным образом, начальное направление выбирается как $s_0 = -\nabla E_0$, а каждый последующий шаг производится в заданном направлении

$$s_{k+1} = -g_{k+1} + s_k \frac{(g_k - g_{k-1}, g_{k+1})}{(g_k - g_{k-1}, s_{k+1})},$$

где $g_k = \nabla E_k$, $k=0,1,...$ – номер итерации. В этом направлении выявляется локальный минимум функции ошибки $E(w)$.

Модель на основе гибридной нейронной сети. Гибридная сеть является каскадным объединением сети Кохонена и двухслойного персептрона [15]. Самоорганизующаяся сеть Кохонена выделяет значимые факторы во входных векторах. Эти факторы образуют входной слой персептрона (число входов персептрона совпадает с количеством кластеров, выделенных сетью Кохонена). Количество выходных нейронов совпадает с количеством значимых характеристик, где каждая характеристика значима хотя бы для одного кластера. Количество скрытых нейронов определяется экспериментально.

Гибридная сеть обрабатывает все кластеры, но при этом требуется дополнительная модификация задачника, необходимость которой вызвана наличием большого числа факторов, несущественных для некоторых кластеров. Модификация задачника состоит в обнулении тех компонент выходных векторов, которые являются малозначимыми для кластера, соответствующего входному вектору.

Недостаток гибридной нейронной сети состоит в сложности ее обучения. Показано [21], что наличие большого числа нулевых компонент в задачнике влечет усложнение обучения градиентными методами, поскольку на поверхности ошибки нейронной сети, являющейся геометрической интерпретацией функции ошибки, возникают обширные плоские области. Использование гибридной сети оправдано при значительном пересечении множеств значимых характеристик между кластерами.

Модель на основе комплекса многослойных нейронных сетей. Для каждого кластера входных векторов, выделенного сетью Кохонена, предлагается использовать отдельный персептрон с одним скрытым слоем. Задачник сети представляет собой множество пар «вход–выход» исходного задачника, где вход принадлежит исследуемому кластеру. Количество входных нейронов персептрона совпадает с количеством нейронов сети Кохонена, количество скрытых нейронов определяется экспериментально, количество выходных нейронов определяется числом значимых факторов исследуемого кластера.

Данная модель позволяет решить проблему нулевых значений в задачнике для гибридной нейронной сети, однако более громоздка, поскольку для каждого кластера создается отдельная аппроксимирующая нейронная сеть. Такая модель оправдана при небольшом количестве значимых характеристик для кластеров или небольшом количестве кластеров.

Алгоритм обучения нейросетевых моделей имеет ряд особенностей:

1. Качество результата каждого этапа (кластерный анализ, факторный анализ, регрессионный анализ) оказывает ключевое влияние на следующий этап.
2. Качество обучения не ухудшается, если число выходных нейронов сети Кохонена превысит количество кластеров входных векторов. Это объясняется тем, что разбиение множества схожих векторов на более мелкие кластеры не влияет негативно на проведение этапа факторного анализа.
3. Если в кластер попадает большая доля неверно распознанных векторов, при использовании модели на основе гибридной нейронной сети значение ошибки в проблемном кластере снижается, что связано с малым количеством значимых характеристик в этом кластере.

Описанные выше особенности позволяют выбрать модель регрессионного анализа на основе результатов кластерного и факторного анализа:

1. В случае выделения большого числа кластеров входных векторов предпочтительнее использование модели на основе гибридной нейронной сети.
2. В случае слабых пересечений наборов значимых характеристик между кластерами предпочтительнее использование модели на основе комплекса многослойных персептронов.
3. В случае выделения большого числа кластеров со слабыми пересечениями наборов значимых характеристик возможно использование модели на основе комплекса гибридных нейронных сетей, в которой каждая гибридная сеть обрабатывает кластеры с сильными пересечениями наборов значимых характеристик.

Тестирование алгоритма

В качестве исходного алгоритма ранжирования рассмотрен классический алгоритм OkapiBM25 [20], в котором значение функции релевантности рассчитывается следующим образом:

$$BM25(d, q) = \sum_{t \in q} BM25w_{t,d},$$

где

$$BM25w_{t,d} = idf_t * \frac{(k_1 + 1)tf_{t,d}}{k_1(1 - b + b \frac{dl}{avdl}) + tf_{t,d}},$$

d – документ; q – запрос; t – леммы запроса; dl – длина документа; $avdl$ – средняя длина документов в коллекции; $tf_{t,d}$ – частота леммы t в документе d ; idf_t – обратная частота встречаемости слова t ; $k_1=2$, $b=0,75$.

Обучающая выборка строится на основе текстовой коллекции РОМИП-2003 и запросов из задания РОМИП-2006 [22]. Из задания РОМИП-2006 отобраны запросы, количество слов которых варьируется от 2 до 5, не включают в себя цифры, слова с опечатками, неизвестные слова и при этом в текстовой коллекции содержится не менее 5 документов, включающих все слова запроса. Вектор q состоит из 11 компонент, часть из которых может быть нулевыми. Первые 5 пар компонент содержат значения $tf_{t,d}$ и idf_t , $t \in q$, которые являются факторами, описывающими запрос, последняя компонента определяет количество слов запроса.

Входными векторами задачника являются векторы параметров запросов q , выходными – векторы d параметров документов, получивших 1 ранг при ранжировании по алгоритму OkapiBM25. Вектор d содержит значения $tf_{t,d}$ и dl , длина вектора d равна 6. В задачник вошел 141 запрос длины 2, 158 запросов длины 3, 157 запросов длины 4 и 133 запроса длины 5. 80 % векторов использованы в качестве обучающих, а 20 % – в качестве тестовых.

В качестве функции ошибки использована функция

$$MSE = \frac{1}{MN} \sum_{i=1}^N \sum_{j=1}^M (y_{i,j} - d_{i,j})^2,$$

где N – количество примеров; M – размерность выходных векторов; $y_{i,j}$ – компоненты выходного вектора нейронной сети; $d_{i,j}$ – компоненты ожидаемого выходного вектора.

Количество нейронов сети Кохонена (количество выделяемых сетью кластеров) наращивалось, начиная с двух, до 8, исходя из качества работы гибридной сети или комплекса сетей. В качестве критерия качества использовалась ошибка обучения. Дальнейшее увеличение числа нейронов не приводит к улучшению качества работы сети, но приводит к увеличению объема вычислений.

В ходе предварительной обработки данных исключена компонента, соответствующая количеству слов запроса, поскольку она является линейной комбинацией 5 других компонент. В результате факторного анализа входные векторы разделились на кластеры по количеству слов в запросе. Оптимальное разбиение получено при использовании 8 нейронов в слое Кохонена: 2 кластера соответствовали длине запроса 2, 2 кластера соответствовали длине запроса 3, 2 кластера – длине запроса 4 и 2 кластера длине запроса 5. Для каждого кластера в качестве значимых факторов определены $tf_{t,d}$ и dl , где $t \in q$ для большой доли документов. Размер доли определяется параметрами ε и p . В исследовании использованы значения $\varepsilon=0,01$ и $p=0,25$, что соответствует 75% -й доле. В скрытом слое персептрона гибридной сети использовалось 16 нейронов, в скрытых слоях комплекса персептронов использовалось 8 нейронов.

Таблица 1. Результаты обучения модели на основе гибридной сети

Table 1. Results of model learning based on hybrid network

Номер кластера Cluster no.	Размер кластера (обучающие данные) Cluster size (learning data)	Ошибка на обучающих данных Error on learning data	Доля неверных ответов Part of wrong answers
1	51	0,0001	0
2	63	0,0001	
3	58	0	
4	61		
5	64	0,0533	0,03125
6	50	0,0003	0
7	60	0,0002	0,01667
8	63	0,0001	0

Ответом обученной нейронной сети на входной вектор запроса являются значения компонент вектора d , которые обеспечат высокий ранг документа по данному запросу. Ответ нейронной сети считается неверным, если релевантность документа ниже, чем у векторов-документов задачника.

В табл. 1 и 2 представлены ошибки, полученные на обучающих и тестовых данных (доля неверных ответов), на восьми различных кластерах, выделенных сетью Кохонена. Результаты тестирования алгоритма показывают успешное обучение моделей и низкие значения ошибок обучения и тестирования.

Таблица 2. Результаты обучения модели на основе комплекса многослойных персептронов**Table 2.** Results of model learning based on complex of multilayered perceptrons

Номер кластера Cluster no.	Размер кластера (обучающие данные) Cluster size (learning data)	Ошибка на обучающих данных Error on learning data	Доля неверных ответов Part of wrong answers
1	51	0,000189	0,0589
2	63	0,0001	0,0476
3	58	0,00062	0
4	61	0,0002	
5	64	0,00017	
6	50	0,0001	
7	60	0,0001	0,01667
8	63	0,0001	0,01589

Заклучение

Предложен алгоритм идентификации систем текстового поиска в базах данных систем геомониторинга, включающий факторный и регрессионный анализ данных. Факторный анализ включает кластеризацию данных на основе использования

сети Кохонена. Для проведения регрессионного анализа предлагается использовать одну из двух нейросетевых моделей: на основе гибридной нейронной сети или на основе комплекса многослойных персептронов. Результаты тестирования алгоритма показывают успешное обучение моделей и низкие значения ошибок обучения и тестирования. При этом модель на основе гибридной нейронной сети имеет сложности в обучении при наличии большого числа непересекающихся значимых характеристик.

Наиболее узким местом алгоритма является метод факторного анализа, позволяющий выделить значимые факторы. Кроме предложенного статистического метода, выделяющего наиболее важные факторы, возможно применение сети Байеса [23] или методов нечеткой логики [24]. Следующим шагом в данной разработке может стать анализ влияния сложных факторов, являющихся нелинейной комбинацией измеряемых характеристик. Также возможны модификации моделей введением веса значимости характеристики, влияющего на обучение нейросетевых моделей и интерпретацию результатов.

СПИСОК ЛИТЕРАТУРЫ

1. «ГЕОМОНИТОРИНГ» – Интернет-сайт состояния геологической среды Российской Федерации. URL: <http://www.gisinfo.ru/projects/18.htm> (дата обращения: 26.05.2014).
2. Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval: The Concepts and Technology behind Search. 2nd edn. – USA: Addison Wesley Professional, 2011. – 913 p.
3. Гулин А., Карпович П. Жадные алгоритмы в задачах оптимизации качества ранжирования. – 2009. URL: http://download.yandex.ru/company/experience/GDD/Zadnie_algoritmy_Karpovich.pdf (дата обращения: 13.08.2014).
4. Large scale machine learning systems and methods / J. Bem, G.R. Harik, J.L. Levenberg, N. Shazeer, S. Tong. – 2013. URL: <http://www.google.com/patents/US8688705> (дата обращения: 13.08.2014).
5. Liu T.Y. Learning to rank for information retrieval. – Berlin: Springer, 2011. – 288 p.
6. Pintelon R., Schoukens J. System identification: a frequency domain approach. – Hoboken: John Wiley & Sons, 2012. – 788 p.
7. Ашманов И., Иванов А. Оптимизация и продвижение сайтов в поисковых системах. 3-е изд. – СПб.: Питер, 2011. – 464 с.
8. Дистанционное зондирование Земли. URL: <http://www.spacecorp.ru/directions/sensing/> (дата обращения: 13.08.2014).
9. Зябрев И., Пожарков О. Статистические методы исследования алгоритмов текстового ранжирования поисковых систем. – 2009. URL: <http://www.altertrader.com/publications18.html> (дата обращения: 13.08.2014).
10. Гулин А., Маслов М., Сегалович И. Алгоритм текстового ранжирования Яндекса на РОМИП-2006 // Труды четвертого российского семинара РОМИП'2006. – СПб.: НУ ЦСИ, 2006. – С. 40–51.
11. Зябрев И., Пожарков О., Пожаркова И. Алгоритм отбора максимально эффективного множества доноров для продвижения сайта в поисковых системах. – 2011. URL: <http://www.altertrader.com/publications25.html> (дата обращения: 13.08.2014).
12. Зябрев И., Пожарков О., Пожаркова И. Реализация и анализ эффективности метода построения оптимального множества доноров для продвижения сайта в поисковых системах. – 2011. URL: <http://www.altertrader.com/publications31.html> (дата обращения: 13.08.2014).
13. Зябрев И., Пожарков О., Пожаркова И. Моделирование алгоритма текстового ранжирования Яндекса при помощи Matrix-Net. – 2012. URL: <http://www.altertrader.com/publications21.html> (дата обращения: 13.08.2014).
14. Ежов А.А., Шумский С.А. Нейрокомпьютеринг и его применения в экономике и бизнесе. – М.: ИНТУИТ, БИНОМ, Лаборатория знаний, 2007. – 222 с.
15. Осовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002. – 344 с.
16. Тарков М.С. Нейрокомпьютерные системы. – М.: ИНТУИТ, БИНОМ, Лаборатория знаний, 2006. – 140 с.
17. Bidyadhar S., Debashisha J. A differential evolution based neural network approach to nonlinear system identification // Applied Soft Computing. – 2011. – V. 11. – Iss. 1. – P. 861–871.
18. Self-Organising Maps in Document Classification: a Comparison with Six Machine Learning Methods / J. Saarikoski, J. Laurikkala, K. Järvelin, M. Juhola // Adaptive and Natural Computing Algorithms Lecture Notes in Computer Science. – 2011. – V. 6593. – С. 260–269.
19. Han J., Kamber M., Pei J. Data mining: concepts and techniques. 3rd ed. – Burlington: Elsevier Inc, 2012. – 703 p.
20. Upstill T.G. Document ranking using web evidence. PhD thesis. – Canberra, 2005. – 228 с.
21. Kordos M., Duch W. A survey of factors influencing MLP error surface // Control and Cybernetics. – 2004. – V. 33. – № 4. – P. 611–631.
22. Российский семинар по оценке методов информационного поиска (РОМИП). Текстовая коллекция РОМИП-2003 и запросы из задания РОМИП-2006. URL: <http://romip.ru/> (дата обращения: 13.08.2014).
23. Терехов С.А. Введение в байесовы сети // Нейроинформатика-2003: V Всеросс. науч.-техн. конф. Лекции по нейроинформатике. Ч. 1. – М.: МИФИ, 2003. – С. 149–187.
24. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. – М.: Горячая линия – Телеком, 2004. – 452 с.

Поступила 30.10.2014 г.

UDC 004.89

REGRESSION ANALYSIS OF GEOMONITORING SYSTEMS DATABASE TEXT RANKING ALGORITHM USING NEURAL NETWORKS

Mikhail S. Tarkov,

A.V. Rzhanov Institute of Semiconductor Physics SB RAS,
13, Lavrentiev avenue, Novosibirsk, 630090, Russia. E-mail: tarkov@isp.nsc.ru

Oyuna A. Kozhushko,

Novosibirsk State University, 2, Pirogova Street,
Novosibirsk, 630090, Russia. E-mail: oyuna@mail.ru

The relevance of the discussed issue is caused by the need to investigate the behavior of test ranking algorithms. The practical value of the research consists in searching for engines developers including the solution of problems of recognition and adaptive classification of objects according to satellite geomonitoring systems.

The main aim of the study is to investigate a neural network model of the geomonitoring database text documents ranking algorithm. The model is built on the basis of Kohonen network, multilayer perceptrons, and k-means clustering method.

The methods used in the study: software implementation and testing of the neural network ranking algorithms by comparing their work results with the results of the classical ranking algorithm OkapiBm25.

The results. The authors have proposed the algorithm, built on the basis of the neural network models of data processing and comprising factor and regression analysis, for geomonitoring database text retrieval systems identification. Factor analysis includes data clustering based on the use of Kohonen network. To simplify the learning, the factor analysis algorithm is developed to eliminate the characteristics irrelevant to rank. The result of the models operation is a set of important ranking characteristics and their optimal values. To perform a regression analysis, it is proposed to use one of two neural network models based on a hybrid neural network or a multilayer perceptrons complex. The regression analysis model is selected on the base of the cluster and factor analysis results. In the case of allocating a large number of the input vectors clusters, a neural network hybrid model is preferable. In the case of the weak intersections between the clusters sets of the significant characteristics, a model based on a set of multilayer perceptrons is preferable. The algorithm testing results show the successful models learning and the low training and testing error values. The proposed models are approved on the OkapiBm25 algorithm's test data, and their application peculiarities are identified depending on the input data characteristics.

Key words:

Geomonitoring systems databases, text ranking algorithm, regression analysis, factor analysis, classification, clustering, neural networks, Kohonen network, multilayer perceptron.

REFERENCES

1. GEOMONITORING – Internet-sait sostoyaniya geologicheskoy sredy Rossiyskoy federatsii [GEOMONITORING – Internet Site of the Geological Environment of the Russian Federation]. Available at: <http://www.gisinfo.ru/projects/18.htm> (accessed 26 May 2014).
2. Baeza-Yates R., Ribeiro-Neto B. *Modern Information Retrieval: the Concepts and Technology behind Search*. 2nd ed. USA, Addison Wesley Professional, 2011. 913 p.
3. Gulín A., Karpovich P. *Zhadnye algoritmy v zadachakh optimizatsii kachestva ranzhirovaniya* [Greedy function optimization in learning to rank]. 2009. Available at: http://download.yandex.ru/company/experience/GDD/Zadnye_algoritmy_Karpovich.pdf (accessed 13 August 2014).
4. Bem J., Harik G. R., Levenberg J. L., Shazeer N., Tong S. *Large scale machine learning systems and methods*. 2013. Available at: <http://www.google.com/patents/US8688705> (accessed 13 August 2014).
5. Liu T.Y. *Learning to rank for information retrieval*. Berlin, Springer, 2011. 288 p.
6. Pintelon R., Schoukens J. *System identification: a frequency domain approach*. Hoboken, John Wiley & Sons, 2012. 788 p.
7. Ashmanov I., Ivanov A. *Optimizatsiya i prodvizhenie saytov v poiskovykh sistemakh* [Optimization and Website Promotion in Search Engines]. St-Petersburg, Piter Publ., 2011. 464 p.
8. *Distantsionnoe zondirovanie Zemli* [Remote sensing]. Available at: <http://www.spacecorp.ru/directions/sensing/> (accessed 13 August 2014).
9. Zyabrev I., Pozharkov O. *Statisticheskie metody issledovaniya algoritmov tekstovogo ranzhirovaniya poiskovykh sistem* [Statistical methods for investigating text ranking algorithms of search engines]. 2009. Available at: <http://www.altertrader.com/publications18.html> (accessed 13 August 2014).
10. Gulín A., Maslov M., Segalovich I. *Algoritm tekstovogo ranzhirovaniya Yandeks na ROMIP-2006* [Text ranking algorithm of Yandex on ROMIP-2006]. *Trudy chetvertogo rossiyskogo seminar ROMIP'2006* [Proc. of the fourth Russian seminar ROMIP'2006]. St-Petersburg, Scientific institution «Center for Strategic Studies», 2006. pp. 40–51.
11. Zyabrev I., Pozharkov O., Pozharkova I. *Algoritm otbora maksimalno effektivnogo mnozhestva donorov dlya prodvizheniya saytov v poiskovykh sistemakh* [Algorithm of selection of the most effective set of donors for website promotion in search engines]. 2011. Available at: <http://www.altertrader.com/publications25.html> (accessed 13 August 2014).
12. Zyabrev I., Pozharkov O., Pozharkova I. *Realizatsiya i analiz effektivnosti metoda postroeniya optimalnogo mnozhestva donorov dlya prodvizheniya sayta v poiskovykh sistemakh* [Implementation and analysis of the effectiveness of the method of constructing the optimal set of donors for the website promotion in the search engines]. 2011. Available at: <http://www.altertrader.com/publications31.html> (accessed 13 August 2014).
13. Zyabrev I., Pozharkov O., Pozharkova I. *Modelirovanie algoritma tekstovogo ranzhirovaniya Yandeks pri pomoshchi MatrixNet* [Modeling algorithm of Yandex text ranking using MatrixNet]. 2012. Available at: <http://www.altertrader.com/publications21.html> (accessed 13 August 2014).
14. Ezhov A.A., Shumskiy S.A. *Neyrokomputing i ego primeneniya v ekonomike i biznese* [Neurocomputing and its applications in Economics and Business].

- nomics and business]. Moscow, INTUIT, BINOM. Laboratoriya znaniy Publ., 2007. 222 p.
15. Osovskiy S. *Neyronnye seti dlya obrabotki informatsii* [Neural networks for information processing]. Moscow, Finansy i statistika Publ., 2002. 344 p.
16. Tarkov M.S. *Neyrokomp'yuternye sistemy* [Neurocomputer systems]. Moscow, BINOM. Laboratoriya znaniy Publ., 2006. 140 p.
17. Bidyadhar S., Debashisha J. A differential evolution based neural network approach to nonlinear system identification. *Applied Soft Computing*, 2011, vol. 11, Iss. 1, pp. 861–871.
18. Saarikoski J., Laurikkala J., Järvelin K., Juhola M. Self-Organizing Maps in Document Classification: A Comparison with Six Machine Learning Methods. *Adaptive and Natural Computing Algorithms Lecture Notes in Computer Science*, 2011, vol. 6593, pp. 260–269.
19. Han J., Kamber M., Pei J. *Data mining: concepts and techniques*. 3rd ed. Burlington, Elsevier Inc, 2012. 703 p.
20. Upstill T.G. *Document ranking using web evidence*. PhD thesis. Canberra, 2005. 228 p.
21. Kordos M., Duch W. A survey of factors influencing MLP error surface. *Control and Cybernetics*, 2004, vol. 33, no. 4, pp. 611–631.
22. *Rossiyskiy seminar po otsenke metodov informatsionnogo poiska (ROMIP). Tekstovaya kolleksiya ROMIP-2003 i zaprosy iz zadaniya ROMIP-2006* [Russian Information Retrieval Evaluation Seminar (ROMIP). Text collection ROMIP-2003 and requests from the task of ROMIP-2006]. Available at: <http://romip.ru/> (accessed 13 August 2014).
23. Terekhov S.A. Vvedenie v bayesovy seti [Introduction into Bayesian Networks]. V *Vserossiyskaya nauchno-tehnicheskaya konferentsiya «Neuroinformatika-2003». Lektsii po neuroinformatike. Glava 1* [V Russian Science and Technology Conference. Neuroinformatics-2003. Lectures on neuroinformatics. Ch. 1]. Moscow, MIFI Press, 2003. pp. 149–187.
24. Rutkovskaya D., Pilinskiy M., Rutkovskiy L. *Neyronnye seti, geneticheskie algoritmy i nechetkie sistemy* [Neural networks, genetic algorithms and fuzzy systems]. Moscow, Goryachaya liniya – Telekom Publ., 2004. 452 p.

Received: 30 October 2014.