

ИСПОЛЬЗОВАНИЕ ИЕРАРХИЧЕСКОЙ ВРЕМЕННОЙ ПАМЯТИ ДЛЯ ИДЕНТИФИКАЦИИ СИСТЕМЫ РАНЖИРОВАНИЯ ДОКУМЕНТОВ

О. А. Кожушко, М. С. Тарков*

Новосибирский государственный университет,
630090, Новосибирск, Россия.

Институт физики полупроводников им. А. В. Ржанова СО РАН*,
630090, Новосибирск, Россия.

УДК 004.89

Предложена модель иерархической временной памяти (ИВП) для идентификации системы ранжирования текстовых документов. Предложен подход к выбору параметров модели, и дана оценка времени обучения модели. Тестирование модели проведено на данных алгоритма OkapiBm25, примененного к коллекции текстовых документов семинара РОМИП. Полученные результаты позволяют судить о перспективности модели ИВП для решения поставленной задачи.

Ключевые слова: алгоритм ранжирования, идентификация системы, иерархическая временная память.

A model of hierarchical temporal memory for text documents ranking system identification is proposed. An approach to the model's parameters evaluating is proposed and a training time evaluation is given. Tests were performed on data received by modeling OkapiBm25 algorithm applied to the ROMIP seminar text documents collection. The obtained results allow us to conclude that the model can solve the identification problems.

Key words: system identification, text document ranking algorithm, hierarchical temporal memory.

Введение. В настоящее время большое внимание уделяется задаче поиска информации в коллекциях документов (information retrieval) [1]. Современные поисковые системы постоянно модифицируют свои алгоритмы, подстраиваясь под новые особенности коллекций документов и ожидания пользователей. Одна из наиболее важных подзадач информационного поиска — это задача ранжирования найденных документов по степени их релевантности запросу пользователя. Успешное решение данной задачи позволит пользователям в первую очередь ознакомиться с документами, максимально релевантными их запросу. Поскольку объем данных непрерывно увеличивается, а сами данные видоизменяются, естественным шагом для решения данной задачи стало использование машинного обучения для построения функции релевантности [2]. С одной стороны, данный подход позволяет повысить эффективность поисковой системы, но с другой — превращает алгоритм ранжирования в „черный ящик“, что приводит к проблеме идентификации системы ранжирования и поиска ее ключевых элементов.

Большинство исследователей уделяет внимание качеству поиска, оценивая его эффективность с помощью статистических метрик [1]. Идентификация алгоритма позволяет

оценить поведение алгоритма ранжирования в зависимости от разных входных данных. Примером такого анализа является работа по моделированию алгоритма ранжирования Яндекс [3] с помощью жадного алгоритма построения деревьев решений. Этот метод позволил выявить важные факторы ранжирования, но является эмпирическим и не имеет вероятностного обоснования эффективности.

В данной работе для идентификации алгоритма ранжирования используется иерархическая временная память (ИВП, *hierarchical temporal memory*), построенная на основе теории Дж. Хокинса [4, 5]. Модель ИВП строит иерархическое представление исследуемого объекта, обучаясь на последовательности представляемых ей образов. Для идентификации алгоритма ранжирования это означает возможность учета сложных факторов, являющихся комбинацией простых характеристик. Сходство этой модели с байесовыми сетями позволит выбрать значимые при ранжировании факторы на основе статистической информации.

1. Постановка задачи. Основным механизмом работы алгоритма ранжирования поисковой системы задает функция релевантности f , которая сопоставляет паре векторов $\langle q, d \rangle$, описывающих текстовый запрос q и документ d соответственно, числовую оценку релевантности

$$f : (q, d) \rightarrow r.$$

Найденные документы сортируются по убыванию значения функции релевантности. К значению функции релевантности могут быть применены фильтры, снижающие итоговое значение релевантности документов, релевантность которых была искусственно завышена.

Алгоритм ранжирования осуществляет функцию вида

$$F_D = (q, d) \rightarrow \text{rank}_D(f(q, d)),$$

где D — рассматриваемая коллекция документов, а функция rank сопоставляет документу порядковый номер в списке документов коллекции, отсортированном по убыванию значения функции релевантности.

Задача идентификации системы ранжирования является дуальной к задаче вычисления релевантности найденных документов запросу и подразумевает построение модели, устанавливающей взаимосвязь между входными и выходными значениями данной системы. В данной статье задача идентификации ставится в виде задачи классификации.

Пусть определено M классов релевантности, а функция $\text{class}(\text{rank}_D(f(q, d)))$ задает номер класса релевантности по рангу, присвоенному алгоритмом ранжирования. Необходимо построить идентифицирующую модель M_f , такую, что на заданном множестве примеров $X = \{\langle q, d \rangle_i \in R^m, i = 1, \dots, N\}$,

$$E(f, F_D, X) = \frac{1}{|X|} \sum_{i=1}^N I(\text{class}(F_D(\langle q, d \rangle_i)), \text{class}(M_f(\langle q, d \rangle_i))) < \varepsilon,$$

где E — функция ошибки, ε — заданная константа,

$$I(x_1, x_2) = \begin{cases} 1, & \text{если } x_1 = x_2, \\ 0, & \text{иначе.} \end{cases}$$

Указанная функция ошибки определяет долю пар $\langle q, d \rangle$, неверно классифицированных по степени релевантности. Таким образом, требуется построить модель, которая на достаточной доле тестовых примеров присваивает ту же степень релевантности парам $\langle q, d \rangle$, что и исходный алгоритм.

2. Иерархическая временная память. Иерархическая временная память представляет собой обучающуюся модель с многослойной древовидной структурой. Она состоит из узлов, которые объединяются в слои. Все слои нумеруются в направлении от нижнего слоя к верхнему. Слои делят на три типа: входной слой, обрабатывающий входную информацию, промежуточные слои, в которых проводятся промежуточные расчеты, и выходной слой, генерирующий результат работы ИВП. Узлы каждого слоя не связаны между собой. Узел входного слоя имеет связь с одним узлом первого промежуточного слоя. Узел промежуточного слоя связан с несколькими узлами предыдущего слоя и одним узлом следующего слоя, узел выходного слоя имеет связь со всеми узлами последнего промежуточного уровня.

Функционирование каждого узла сети состоит из двух этапов, называемых пространственным объединением и временной группировкой. В ходе пространственного объединения узел разделяет входные данные на группы похожих векторов, где сходство определяется с помощью заданных функции расстояния и порогового значения. В ходе временной группировки узел находит часто встречающиеся временные последовательности во входных данных, которые указывают на исходную зависимость. Иерархическая временная память и ее узлы могут функционировать в двух режимах: обучения и тестирования.

2.1. *Обучение ИВП.* Слои ИВП обучаются последовательно от нижнего слоя к верхнему, при этом узлы каждого слоя обучаются параллельно. Входной сигнал промежуточного или выходного слоя формируется с помощью конкатенации выходных сигналов узлов предыдущего слоя или компонент входных векторов. Каждый узел, просматривая входную последовательность данных, проводит пространственную группировку, вычисляет матрицу смежности, затем формирует временные группы. При пространственной группировке вычисляются пространственные центры. Текущий входной вектор фиксируется как пространственный центр в том случае, если ранее не был зафиксирован достаточно близкий к нему пространственный центр. Для входного узла сходство входного вектора и центра определяется как расстояние между ними в евклидовом пространстве. Для центра промежуточного и выходного узлов входной вектор представляет собой вектор индексов временных групп предыдущего слоя, и сходство между векторами определяется функцией

$$dist(x, c) = \sum_{i=1}^n I(x_i, c_i),$$

где n — длина векторов x и c . Входной вектор x образует новый пространственный центр, если $dist(x, c) > mdist$ для всех пространственных центров c , где $mdist$ — заданный порог.

В ходе временной группировки входного и промежуточного узла пространственные центры $\{c^i, i = 1, \dots, n_c\}$ объединяются во временные группы $\{g^j, j = 1, \dots, n_g\}$. Временная группа строится как цепь Маркова с матрицей смежности T . Компоненты T_{ij} равны количеству следований центра c_i за центром c_j . Длина группы ограничена заданной величиной $groupMaxSize$. При завершении обучения узла формируется матрица PCG , компоненты которой равны условной вероятности $PCG_{ij} = P(c_i | g_j)$ появления центра c_i при условии появления группы g_j .

Выходной узел обучается с учителем. Вместо формирования временных групп выходной узел вычисляет условную вероятность $PCW_{ij} = P(c_i|w_j)$ появления пространственного центра c_i при принадлежности заданному классу w_j . Далее по теореме Байеса подсчитывается вероятность появления класса w_j :

$$P(w_j) = \frac{\sum_{i=1}^{n_c} PCW_{ij}}{\sum_{i=1}^{n_c} \sum_{j=1}^{n_l} PCW_{ij}},$$

где n_l — количество распознаваемых классов.

2.2. *Тестирование ИВП.* В режиме тестирования ИВП входной сигнал λ^- идет от нижнего уровня к верхнему. Функционирование каждого узла сводится к сопоставлению сигналу λ^- максимально правдоподобной временной группы g^j или класса w_i . Входной узел генерирует вектор y вероятности совпадения λ^- с пространственными центрами:

$$y_i = \exp\left(-\frac{\|c^i - \lambda^-\|^2}{\sigma^2}\right).$$

Промежуточный и выходной узлы генерируют вектор y как $y_i = \prod_{j=1}^k \lambda_j^-[c_j^i]$, где $\lambda_j^-[c_j^i]$ — вероятность совпадения входного сигнала с наиболее вероятной временной группой j -го дочернего узла, а k — количество дочерних узлов. Входной и промежуточный узлы вычисляют апостериорную вероятность появления сигнала λ^- при условии появления группы g^j

$$\lambda_j^+ = P(\lambda^-|g^j) = \sum_{i=1}^N y_i \cdot PCG_{ij}.$$

В качестве выходного сигнала выбирается группа $g^{j_{max}}$ с максимальным значением $\lambda_{j_{max}}^+$.

Для выходного узла аналогичным образом подсчитываются

$$y_i = \prod_{j=1}^k \lambda_j^-[c_j^i],$$

$$P(\lambda^-|w_j) = \sum_{i=1}^n y_i \cdot PCW_{ij}.$$

Наконец, выходной сигнал формируется как

$$P(w_j|\lambda^-) = \frac{P(\lambda^-|w_j) \cdot P(w_j)}{\sum_{i=1}^k P(\lambda^-|w_i) \cdot P(w_i)}$$

и равен вероятности класса w_j при поступлении входного сигнала λ^- .

3. Построение модели идентификации. Разработка модели ИВП производится с помощью априори известных данных об исследуемом алгоритме ранжирования. В качестве тестового алгоритма ранжирования в работе рассмотрен классический алгоритм ОкариВМ25 [6]. Алгоритм имеет аддитивную функцию релевантности:

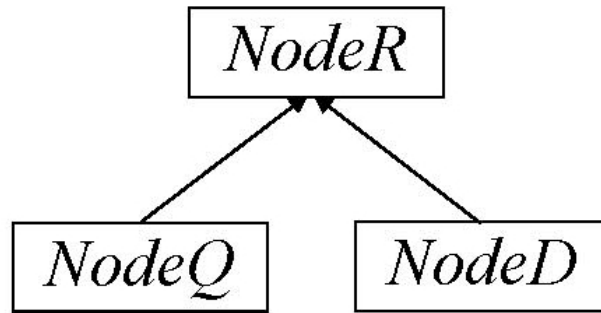


Рис. 1. Модель ИВП для идентификации алгоритма ранжирования документов

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

где

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

где 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 леммы t . Коэффициенты обычно принимаются равными следующим значениям: $k_1 = 2$, $b = 0,75$.

На вход ИВП поступают данные о запросах и документах, ранжируемых по данным запросам, на выход — результат ранжирования. Результат ранжирования рассматривается как степень релевантности запроса документу. В данной работе различается три степени релевантности: „высокая“, „средняя“, „низкая“. Далее описывается модель идентификации алгоритма ранжирования OkapiBm25.

Предлагаемая модель (рис. 1) имеет 2 слоя: входной и выходной. Входной слой содержит 2 узла, один из которых работает с параметрами, описывающими запрос, а второй — с параметрами, описывающими документ. В данном случае входной вектор разбивается на вектор запроса q и вектор документа d , каждый вектор обрабатывается своим узлом.

Входные узлы $NodeQ$ и $NodeD$ обрабатывают входные векторы q и d : выделяют центры кластеров схожих между собой векторов и строят устойчивые группы центров. Выходной узел $NodeR$ получает на вход вектор из двух компонент (x,y) , x — номер группы вектора-запроса q , y — номер группы вектора-документа d .

Первым этапом в обучении ИВП является подготовка обучающей последовательности, состоящей из троек векторов $\langle q^i, d^i, r^i \rangle$. Вектор q^i характеризует i -й запрос, d^i — i -й документ, значение $r^i = class(rank(BM25(q,d)))$ соответствует степени релевантности документа d^i , полученного по запросу q^i . Узлы первого уровня обрабатывают данные разного объема и с различными статистическими характеристиками (такими как выборочное среднее и выборочная дисперсия). При построении обучающей последовательности необходимо „правильно“ отсортировать входные данные. В данной работе сортировка осуществляется следующим жадным алгоритмом [7]:

- 1) Выбрать центр класса x .

2) Пока есть векторы, не вошедшие в упорядоченную последовательность, повторять: выбрать вектор y , ближайший к x ; включить y в последовательность; x присвоить y .

Использование жадного алгоритма также показывает способ определения временных разрывов (temporal gap) [5]. Разрывы обозначают резкие скачки в последовательности данных. Фиксация разрывов необходима для того, чтобы ИВП не запоминала ложные временные группы. В данном случае разрыв возникает при добавлении в последовательность вектора, значительно удаленного от предыдущего.

Таким образом, алгоритм построения обучающей последовательности имеет следующий вид.

1) Входные векторы узла отсортировать жадным алгоритмом, начав с вектора минимальной длины.

2) Обозначить временной разрыв между векторами, если расстояние между ними в последовательности превышает порог gap .

Обучение ИВП по построенной последовательности происходит в три этапа:

1) Параллельное обучение узлов входного уровня.

2) Подготовка обучающей последовательности для выходного узла в виде $\langle q, d, r \rangle$, где q — номер группы-запроса, d — номер группы-документа, r — степень релевантности. Обозначение разрывов не требуется, поскольку выходной узел обучается с учителем.

3) Обучение выходного узла.

Для успешного обучения необходимо правильно выбрать значения параметров ИВП. Для каждого узла необходимо подобрать значения параметров $mdist$, gap , σ^2 , $groupMaxSize$. Исходя из условий задачи, параметр $groupMaxSize$ равен максимальной длине последовательности документов, имеющих одинаковую оценку релевантности по заданному запросу. Параметр σ^2 определяет степень активации пространственного центра в зависимости от удаленности входного вектора и может быть оценен как квадрат средней дисперсии компонент входных векторов [7]. Подбор остальных параметров можно осуществлять с помощью генетического алгоритма, обучая на каждом шаге популяцию систем ИВП [8] или подбирая экспериментально.

После обучения ИВП работает в режиме тестирования, сопоставляя входным векторам q и d степень релевантности документа d запросу q . Кроме этого, данная модель ИВП позволяет получить информацию о векторах документов, которые с наибольшей вероятностью получают определенную оценку релевантности. Для этого необходимо:

1) Вычислить временную группу g^q вектора q .

2) Выбрать все пространственные центры выходного узла $c_{d_i}^q = \langle g^q, g^{d_i} \rangle$, содержащие g^q .

3) Выбрать временную группу g^d , имеющую соответствующую компоненту матрицы PCW в столбце, соответствующем заданному классу высшей степени релевантности. Векторы, входящие в группу g^d , являются искомыми.

4. Результаты экспериментов. Обучающая выборка построена на основе текстовой коллекции РОМИП-2003 и запросов из задания РОМИП-2006 [9]. Отобраны запросы, количество слов в которых варьируется от 2 до 5, при этом запросы не включают в себя цифры, слова с опечатками, неизвестные слова. Всего в задачник вошло 435 запросов, в тестовую выборку — 75 запросов. Для каждого запроса в обучающую и тестовую выборку включено три документа — высоко-, средне- и низкорелевантных. Степень релевантности определялась по рангу: высокорелевантные документы имеют ранг 1, среднерелевант-

Таблица 1

Параметры и результаты обучения ИВП							
<i>mdist</i>		Количество центров		Количество групп		Точность распознавания	
NodeQ	NodeD	NodeQ	NodeD	NodeQ	NodeD	Обучающие данные	Тестовые данные
0,27400	0,38140	8	21	4	6	29,70 %	26,22 %
0,03420	0,09540	367	19	122	9	37,60 %	34,67 %
0,00850	0,00595	422	351	141	120	92,65 %	64,00 %
0,00210	0,00149	428	845	143	284	98,50 %	68,00 %

ные — ранг 11, низкорелевантные — 21. Таким образом, выделено три класса релевантности, принадлежность к которым указывается в качестве выходных значений.

Экспериментально установлено, что малые значения параметров σ^2 узлов приводят к тому, что более 96 % тестовых примеров распознаются как принадлежащие классу под номером 1, независимо от интерпретации данного класса. Это связано с тем, что компоненты выходных векторов $\lambda_j^+ = P(\lambda^- | g^j)$ принимают значения порядка 10^{-40} и меньше, соответственно, выходной узел на большинстве примеров вычисляет нулевой вектор вероятности принадлежности входного вектора классам релевантности и выбирает первый класс. Значимые результаты ИВП дает при значении параметра σ^2 , равного среднеквадратичному отклонению обучающей выборки для данного узла. Дальнейшая настройка параметров проводится экспериментально. Таким образом, в проведенном эксперименте параметр σ^2 оценен как 0,00347 для узла *NodeQ*, обрабатывающего векторы запросов, и 0,00001663 для узла *NodeD*, обрабатывающего векторы документов.

Значения параметров *gap* и *mdist* получены экспериментально. При уменьшении значения параметра *gap* уменьшается размер временных групп. При достижении значений порядка 10^{-2} начинают преобладать одноэлементные группы. При уменьшении параметра *mdist* наблюдается снижение значения ошибки на обучающих и тестовых данных. В таблице приведено количество выделяемых в узлах пространственных центров и временных групп, точность распознавания модели в зависимости от параметра *mdist*.

Время обучения модели на 1305 примерах составляет 47 минут и 54 секунды при тактовой частоте процессора 1,6 Гц. Время обучения системы зависит как от количества примеров в выборке и архитектуры сети, так и от настроек параметров модели, влияющих на количество выделяемых пространственных центров. Время обучения одного узла оценивается как $O(N \cdot K + K^3)$, где N — количество примеров в обучающей выборке, а K — количество пространственных центров узла. Данная оценка напрямую следует из алгоритма обучения.

Заключение. Сформулирована задача идентификации алгоритма ранжирования в терминах задачи классификации. Для решения данной задачи предложена модель иерархической временной памяти для идентификации алгоритма ранжирования, а также предложен подход к выбору параметров модели и дана оценка времени обучения модели. Предложенная модель может быть использована как для получения степени релевантности запросу документа, так и для выявления характеристик документов, принадлежащих к заданному классу релевантности. Полученные результаты свидетельствуют о перспективности развития данного подхода. Дальнейшее повышение точности возможно за счет увеличения обучающей выборки и увеличения количества узлов и слоев в модели ИВП.

Список литературы

1. Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval: The Concepts and Technology Behind Search. Addison Wesley Professional, USA, 2011.
2. Гулин А., Карпович П. Жадные алгоритмы в задачах оптимизации качества ранжирования [Электронный ресурс]. 2009. Режим доступа: http://download.yandex.ru/company/experience/GDD/Zadnie_algoritmy_Karpovich.pdf. — 21.01.2015.
3. Зябрев И., Пожарков О., Пожаркова И. Моделирование алгоритма текстового ранжирования Яндекса при помощи MatrixNet [Электронный ресурс]. 2010. Режим доступа: <http://www.altertraider.com/publications21.htm>. — 21.01.2015.
4. Хокинс Дж., Блексли С. Об интеллекте. М.: Вильямс, 2007.
5. Maltoni D. Pattern Recognition by Hierarchical Temporal Memory // DEIS Technical Report. [Электронный ресурс]. 2011. Режим доступа: http://bias.csr.unibo.it/maltoni/HTM_TR_v1.pdf. — 21.01.2015.
6. Upstill T. Document ranking using web evidence. PhD Thesis. The Australian National University, 2005.
7. Kostavelis I., Gasteratos A. On the optimization of hierarchical temporal memory // Pattern Recognition Letters. 2012. V. 33. N 5. P. 670–676.
8. Болотова Ю. А., Спицын В. Г., Фомин А. Э. Применение модели иерархической временной памяти в распознавании изображений // Известия Томского политехнического университета. 2011. V. 318. N 5. P. 60–63.
9. Российский семинар по оценке методов информационного поиска [Электронный ресурс]. — Режим доступа: <http://romip.ru/> — 21.01.2015

*Кожушко Оюна Алексеевна — аспирант
Новосибирского государственного университета,
e-mail: oyuна@mail.ru, тел. 8-913-764-72-65*

*Тарков Михаил Сергеевич — канд. тех. наук,
Институт физики полупроводников
им. А. В. Ржанова СО РАН,
e-mail: tarkov@isp.nsc.ru.*

Дата поступления — 18.02.2015