Varfolomeeva A.A., Strijov V.V. An algorithm for bibliographic records parsing using structure learning methods // Information Technologies, 2014, 7 : 11-15.

Варфоломеева А.А., Стрижов В.В. Алгоритм разметки библиографических списков методами структурного обучения // Информационные технологии, 2014, 7 : 11-15.

УДК 519.256

А. А. Варфоломеева, студент, Московский физико-технический институт
В. В. Стрижов, к.ф.-м.н., Вычислительный Центр РАН*

Алгоритм разметки библиографических списков методами структурного обучения[1]

Аннотация. В предлагаемой работе решается прикладная задача сегментации структурированных текстов: для каждого сегмента библиографической записи определяется его тип поля в формате BibTeX. Также для каждой записи определяется тип ее библиографического описания.

Такая задача возникает в связи с наличием различных стандартов составления библиографических записей: требуется предложить алгоритм определения типов полей библиографических записей, не зависящий от конкретного стандарта их составления.

Для решения задачи определения типа поля в работе предложен метод составления матриц «объектов» и матриц «ответов» – примеров правильной сегментации. В работе предлагается алгоритм разметки библиографических списков методом структурной регрессии, при этом решается задача выбора параметров регрессионной модели. По результатам сегментирования полей записи с помощью кластеризации определяется тип ее библиорафического описания. Качество полученной модели исследуется на наборе неформатированных библиографических списков. В работе показано, что предлагаемый алгоритм имеет хорошее качество сегментации и кластеризации при наличии достаточной обучающей выборки.

Ключевые слова: разметка текстов, структурное обучение, структурная регрессия, сегментирование, выбор признаков, кластеризация.

A.A. Varfolomeeva, V.V. Strijov

An algorithm for bibliographic records parsing using structure learning methods

Abstract. The paper solves the application problem of structured texts segmentation, namely each segment of a bibliographic record must correspond to its filed type of the BibTeX format and each record must correspond to its bibliographic type.

This problem arises due to the existence of different standards for bibliographic records: an algorithm for determining the types of fields of bibliographic records, which is independent of the specific standards of their composition, should be proposed.

To solve the problem of determining the field type the method of constructing

* E-mail: strijov@ccas.ru

matrix “objects” and matrices “answers” is proposed. The authors offer an algorithm of a bibliography lists parsing using the structure regression method, and the optimization problem of regression model’s parameters is also solved. According to the results of fields' segmentation bibliographic types of the records are clustered. The quality of the constructed model is investigated using a collection of non-parsed bibliography lists. In the paper it is shown the proposed algorithm has good quality of segmentation and clustering, if it has sufficient training sample.

Keywords: text parsing, structure learning, structure regression, segmentation, features selection, clustering.

Введение

Работа посвящена построению модели, прогнозирующей структуру текстовой строки. Решается прикладная задача представления неформатированных библиографических записей в виде структуры в формате BibTeX [1] – стандарта управления коллекцией библиографических записей. Требуется разметить библиографические записи: определить соответствия между текстовыми сегментами и полями записей BibTeX, а также указать тип библиографической записи.

Необходимость форматирования текстовой строки вызвана наличием различных стандартов (ГОСТ 7.82-2001, MLA и др.), определяющих различный порядок следования полей библиографической записи. Кроме того, запись может быть составлена с нарушением стандартов.

Методы, предложенные ранее для решения задачи сегментрирования текстов описаны в [2], где граф цитирований строится с помощью полученной разметки библиографического списка. В [3] описана постановка и решение задачи разметки адресной строки. Используется скрытая марковская цепь, недостатком которой является неточное описание структурных зависимостей внутри исходных данных. Автоматическая разметка библиографических записей представлена в [4]. Для обучения модели которой использовалась библиотека одного из стандартов, в силу чего модель неустойчиво работает на других стандартах.

В данной работе для поиска структуры текстовой строки предлагается использовать методы структурного обучения, описанные в [5, 6, 7]. Эти методы используются в области анализа текстов для определения синтаксических зависимостей в предложении [8]. В предлагаемом алгоритме сегментация записей выполняется на базе метода структурной регрессии [9].

Ставится и решается задача определения соответствий между текстовыми сегментами библиографической записи и набором полей структуры BibTeX.

В табл. 1 приведен пример неформатированной библиографической записи и верные либо неверные соотношения между сегментами записи и набором полей

BibTeX.

В работе решается задача выбора оптимального набора признаков модели. Для этого используется модификация ранее предложенного авторами [10, 11] алгоритма последовательного добавления и удаления признаков. Веса признаков оцениваются для случая логистической регресии [12].

После построения соотношений между текстовыми сегментами и набором полей структуры BibTeX требуется определить тип библиографической записи. Для этого строится новое признаковое описание библиографической записи как подмножество полей структуры BibTeX, присутствующих в библиографической записи. С помощью кластеризации k-means [13] определяется тип каждой записи.

Таким образом, для решения задачи определения типа библиографической записи в первую очередь решается подзадача определения для каждого сегмента типа поля библиографической записи в структуре BibTeX, по результатам которой проводится требуемая кластеризация записей. Работа предложенного метода демонстрируется на наборе неформатированных библиографических списков.

Таблица 1: Пример верного и неверного сегментирования библиографической записи

Kwok T. Y., Yeung D. Y. Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems // IEEE Transactions on Neural Networks, 1997. Vol. 8. Pp. 630–645.		
	Верно	Неверно
Type	Article	Book
Author	*Kwok T.Y., Yeung D. Y.*	*Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems*
Title	*Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems*	*IEEE Transactions on Neural Networks*
Journal	*IEEE Transactions on Neural Networks*	-
Pages	630-645	-
Volume	8	1997
Number	-	8
Year	1997	630-645
Address	-	-
Publisher	-	-
Editor	-	*Kwok T. Y., Yeung D. Y.*
URL	-	-

Постановка задачи

Дан набор из m строк $\{t_1, t_2, \ldots, t_m\}$ – библиографических записей. Каждая запись t состоит из текстовых сегментов $t_i = \{t_i^1, t_i^2, \ldots, t_i^n\}$. Задан набор G порождающих функций $\mathsf{G} = \{g\}$, отображающих j-ый текстовый сегмент i-ой строки s_i^j в вектор-строку признаков $\mathbf{x}_{ij}$, $g : s_i^j \mapsto \mathbf{x}_{ij}$.

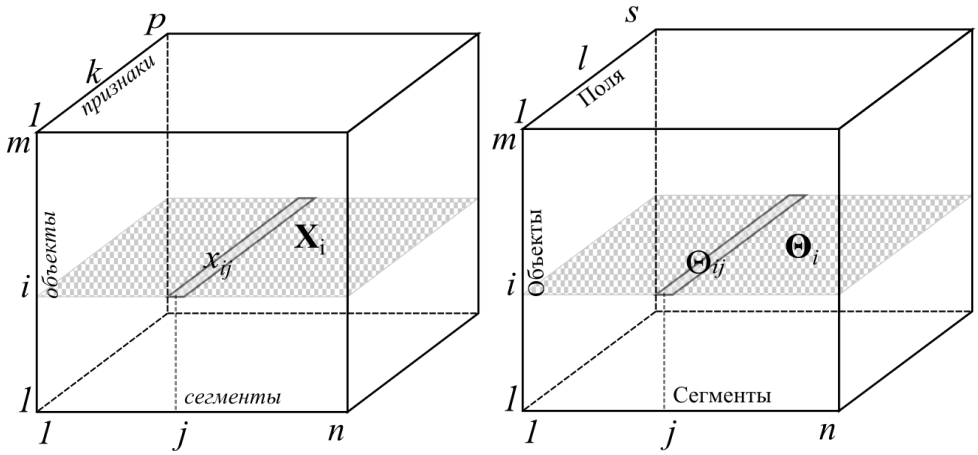


Рис. 1. Вид матриц $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Theta}$

Задана трехиндексная матрица $\mathbf{X}$ «объект – сегмент – признак» размера $m \times n \times p$, где m – число библиографических записей, n – число сегментов, p – число признаков. Каждому объекту, заданному двухиндексной матрицей $\mathbf{X}_i$, где i – номер объекта, поставлена в соответствие бинарная матрица ответов $\mathbf{Y}_i$ размера $n \times s$, где s – число типов полей структуры BibTeX. Элементы матрицы $\mathbf{Y}_i$ отвечают за принадлежность j-го сегмента к ℓ-му типу поля библиографической записи:

$$Y_i(j, \ell) = \begin{cases} 1, \text{если } \mathbf{x}_{ij} \text{ принадлежит к } \ell - \text{ому типу поля;} \\ 0, \text{иначе,} \end{cases} \qquad (1)$$

где $j \in \{1,2,\ldots,n\}$ – индекс сегмента текстовой строки, $\ell \in \{1,2,\ldots,s\}$ – индекс типа поля. Вводится двухиндексная матрица весовых параметров $\mathbf{W}$ размером $p \times s$, элементы которой $w_{k\ell}$ отвечают за значимость k-ого признака для ℓ-ого типа поля, $k = 1,\ldots,p$, $\ell = 1,\ldots,s$. При умножении матрицы $\mathbf{W}$ справа на вектор-строку $\mathbf{x}_{ij}$ признаков j-ого сегмента i-ого объекта получается вектор-строка $\mathbf{\theta}_{ij}$, элементы которой определяют оценку принадлежности данного сегмента к полям структуры BibTeX:

$$\underset{(\ell \times s)}{\mathbf{\theta}_{ij}} = \underset{(1 \times p)(p \times s)}{\mathbf{x}_{ij}\mathbf{W}} . \qquad (2)$$

Тогда оптимальный прогнозируемый тип поля с индексом $\hat{\ell}$ для

признакового описания $\mathbf{x}_{ij}$ сегмента библиографической записи с фиксированным номером i определяется как индекс максимального элемента вектор-строки $\boldsymbol{\theta}_{ij}$:

$$\hat{\ell}_j = \underset{l=1,2,..,s}{\operatorname{argmax}}\, \boldsymbol{\theta}_{ij}(\ell).$$

Аналогично записывая строки $\boldsymbol{\theta}_{ij}$ для каждого вектора признаков $\mathbf{x}_{ij}$ с индексом j объекта $\mathbf{X}_i$, составляется матрица оценок $\boldsymbol{\Theta}_i$, значения которой определяют тип поля каждого сегмента объекта $\mathbf{X}_i$:

$$\hat{Y}_i(j,\ell) = \begin{cases} 1, \text{если } \ell = \hat{\ell}_j; \\ 0, \text{иначе.} \end{cases} \qquad (3)$$

Требуется найти такой набор признаков A из множества G и веса $\mathbf{W}$ этих признаков, что расстояние $\mathrm{Dist}(\hat{\mathbf{Y}}, \mathbf{Y})$ между матрицей ответов $\mathbf{Y}$ и прогнозируемой матрицей $\hat{\mathbf{Y}}$ минимально:

$$\hat{\mathbf{W}} = \underset{\mathbf{W},\mathsf{G}}{\operatorname{argmin}}\, \mathrm{Dist}(\hat{\mathbf{Y}}, \mathbf{Y}) = \underset{\mathbf{W},\mathsf{G}}{\operatorname{argmin}}\, \frac{1}{2} \sum_{i,j,k=1}^{m,n,p} |\hat{Y}_i(j,k) - Y_i(j,k)|. \qquad (4)$$

Искомая матрица весов $\hat{\mathbf{W}}$ определяется минимумом аппроксимированного эмпирического риска Q для случая логистической регрессии.

Введем обозначения, необходимые для определения функции эмпирического риска Q. Матрица весов признаков $\mathbf{W}$ разбивается на s независимых вектор-столбцов $\mathbf{w}_\ell$, соответствующих типу поля ℓ, $\ell \in \{1,2,\ldots,s\}$, и для каждого столбца $\mathbf{w}_\ell$ векторизуется соответствующая ему часть матрицы $\mathbf{Y}$:

$$\underset{(mn\times 1)}{\mathbf{u}_\ell} = \begin{pmatrix} \mathbf{Y}_1(\ell) \\ \vdots \\ \mathbf{Y}_i(\ell) \\ \vdots \\ \mathbf{Y}_m(\ell) \end{pmatrix}, \text{ где } \mathbf{Y}_i(\ell) - \ell - \text{ый столбец матрицы } \mathbf{Y}_i.$$

Матрица $\mathbf{X}$ записывается в двумерном виде:

$$\underset{(mn\times p)}{\mathbf{Z}} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_m \end{pmatrix}, \text{ где } \mathbf{X}_i - \text{матрица признаков } i - \text{го объекта библиографической записи.}$$

Обозначим $\mathbf{z}_i$ – i-ую строку матрицы $\mathbf{Z}$. Тогда функция эмпирического риска записывается в виде

$$Q = \sum_{i=1}^{m\times n} \sum_{\ell=1}^{s} \mathsf{L}(\mathbf{w}_\ell, \mathbf{z}_i, u_\ell(i), \mathsf{A}),$$

где функция потерь

$$\mathsf{L}(\mathbf{w}_\ell, \mathbf{z}_i, u_\ell(i), \mathsf{A}) = \log\bigl(1 + \exp(-\langle \mathbf{w}_\ell{}^{\mathsf{A}}, \mathbf{z}_i{}^{\mathsf{A}} \rangle u_\ell(i))\bigr).$$

При этом $\mathbf{w}_l{}^{\mathsf{A}}$ и $\mathbf{z}_i{}^{\mathsf{A}}$ содержат только подмножество $\mathsf{A} \subset \mathsf{G}$ своих элементов индексов признаков:

$$\hat{\mathsf{A}} = \operatorname*{argmax}_{\mathsf{A} \subset \mathsf{G}} Q(\mathbf{W}, \mathbf{Z}, \mathbf{u}, \mathsf{A}),$$

где $\hat{\mathsf{A}}$ – решение оптимизационной задачи.

Описание алгоритма выбора признаков

Требуется решить задачу выбора оптимального подмножества индексов признаков $\mathsf{A} \subset \mathsf{G}$ и оценки $\hat{\mathbf{W}}(\mathsf{A})$ матрицы весов признаков (4). Для этого используется следующая процедура последовательного добавления и удаления признаков.

Зададим начальное множество признаков $\mathsf{A}_0 = \varnothing$, параметр останова процедуры d и начальные значения функции эмпирического риска $\hat{Q} = Q(\varnothing)$ и номера итерации $t = 0$.

1. Пока мощность набора признаков $|\mathsf{A}_t|$ меньше общего числа признаков p выполняются следующие действия:

(a) увеличиваем номер итерации $t = t + 1$;

(b) находим оптимальный для добавления признак с индексом

$$\hat{g} = \operatorname*{argmin}_{g \in \mathsf{G} \setminus \mathsf{A}_{t-1}} Q(\mathsf{A}_{t-1} \bigcup \{g\});$$

и добавляем его к набору: $\mathsf{A}_t = \mathsf{A}_{t-1} \bigcup \{\hat{g}\}$;

(c) если $Q(\mathsf{A}_t) < \hat{Q}$, то текущее минимальное значение эмпирического риска $\hat{Q} = Q$, номер оптимальной итерации $\hat{t} = t$;

(d) если значение функционала не улучшалось на протяжении d шагов $t - \hat{t} \geq d$, то прервать цикл.

2. Пока мощность набора признаков $|\mathsf{A}_t|$ ненулевая выполняются следующие действия:

(a) увеличиваем номер итерации $t = t + 1$;

(b) находим оптимальный для удаления признак

$$\hat{g} = \operatorname*{argmin}_{g \in \mathsf{A}_{t-1}} Q(\mathsf{A}_{t-1} \setminus \{g\});$$

и удаляем его из набора: $\mathsf{A}_t = \mathsf{A}_{t-1} \setminus \{\hat{g}\}$;

(c) если $Q(\mathsf{A}_t) < \hat{Q}$, то $\hat{Q} = Q$, $\hat{t} = t$;

(d) если значение функционала не улучшалось на протяжении d шагов $t - \hat{t} \geq d$, то прервать цикл.

3. Повторять шаги 1. и 2. пока значения $Q(\mathsf{A}_{\hat{t}})$ убывают.

Алгоритм выбора признаков определяет их оптимальный набор $\hat{\mathsf{A}} = \mathsf{A}_{\hat{t}}$ с

одновременным оцениванием матрицы весов $\mathbf{W}(\hat{\mathrm{A}})$.

Определение типа библиографической записи

Матрица $\hat{\mathbf{Y}}_i$ (3) содержит полную информацию о типах полей ℓ, содержащихся в i-ой библиографической записи. Для решения подзадачи об определении типа записи BibTeX составляется матрица $\mathbf{B}$ размера $m \times s$ по правилу

$$B(i,\ell) = \begin{cases} 1, \text{ если } \ell - \text{ая строка матрицы } \hat{\mathbf{Y}}_i \text{ ненулевая;} \\ 0, \text{ иначе.} \end{cases} \quad (5)$$

Таким образом, элемент $B(i,\ell)$ матрицы $\mathbf{B}$ определяет присутствие в i-ой библиографической записи ℓ-ого типа поля BibTeX, а строка $\mathbf{b}_i$ является новым признаковым описанием объекта – библиографической записи. Поставим задачу разбиения объектов на фиксированное число r кластеров – типов записи в структуре BibTeX.

Обозначим $k(i)$ – номер кластера, к которому отнесена i-ая библиографическая запись и введем следующие функции качества: минимизацию среднего внутрикластерного расстояния

$$F_0 = \frac{\sum_{i<j} [k(i) = k(j)] \rho(\mathbf{b}_i, \mathbf{b}_j)}{\sum_{i<j} [k(i) = k(j)]} \to \min,$$

и максимизацию среднего межкластерного расстояния

$$F_1 = \frac{\sum_{i<j} [k(i) \neq k(j)] \rho(\mathbf{b}_i, \mathbf{b}_j)}{\sum_{i<j} [k(i) \neq k(j)]} \to \max,$$

где индикаторная функция

$$[k(i) = k(j)] = \begin{cases} 1, \text{ если } k(i) = k(j); \\ 0, \text{ иначе.} \end{cases}$$

Задача кластеризации сводится к минимизации функции качества F:

$$F = \frac{F_0}{F_1} \to \min.$$

В качестве метрики $\rho(\mathbf{b}_i, \mathbf{b}_j)$. используется диагонально взвешенная евклидова метрика:

$$\rho(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \Lambda^2 (\mathbf{x}-\mathbf{y})}, \text{ где } \Lambda = \mathrm{diag}(\lambda). \quad (6)$$

Диагональная матрица Λ задает веса, соответствующие признакам описания библиографической записи. Ее значения определяются частотой вхождения поля в библиографические записи.

Кластеризация выполняется с помощью метода k-means. Метод состоит из двух основных шагов:

1. Для каждого элемента $\mathbf{b}_i$ находится ближайший к нему центр кластеров, к которому и относится данный элемент:

$$k(i) = \underset{j=1,\ldots,r}{\operatorname{argmin}} \rho(\mathbf{b}_i, \boldsymbol{\mu}_j).$$

2. Положение центра кластера $\boldsymbol{\mu}_j$ есть центр масс объектов, принадлежащих кластеру:

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{|D_j|} [k(i) = j]\mathbf{b}_i}{\sum_{i=1}^{|D_j|} [k(i) = j]}, \text{ где } |\mathrm{D}_j| - \text{мощность } j - \text{го кластера., j} \in \{1,\ldots,\mathrm{r}\},$$

где r – заданное число кластеров.

Шаги алгоритма повторяются, пока кластеризация объектов меняется.

Начальное приближение центров кластеров $\boldsymbol{\mu}_j$, $j = 1,\ldots,r$ считается заданным. Результатом работы алгоритма является вектор-столбец $\mathbf{k}$, элементы которого $k(i) = j$ отвечают за принадлежность i-ой библиографической записи к j-ому типу записи.

Вычислительный эксперимент

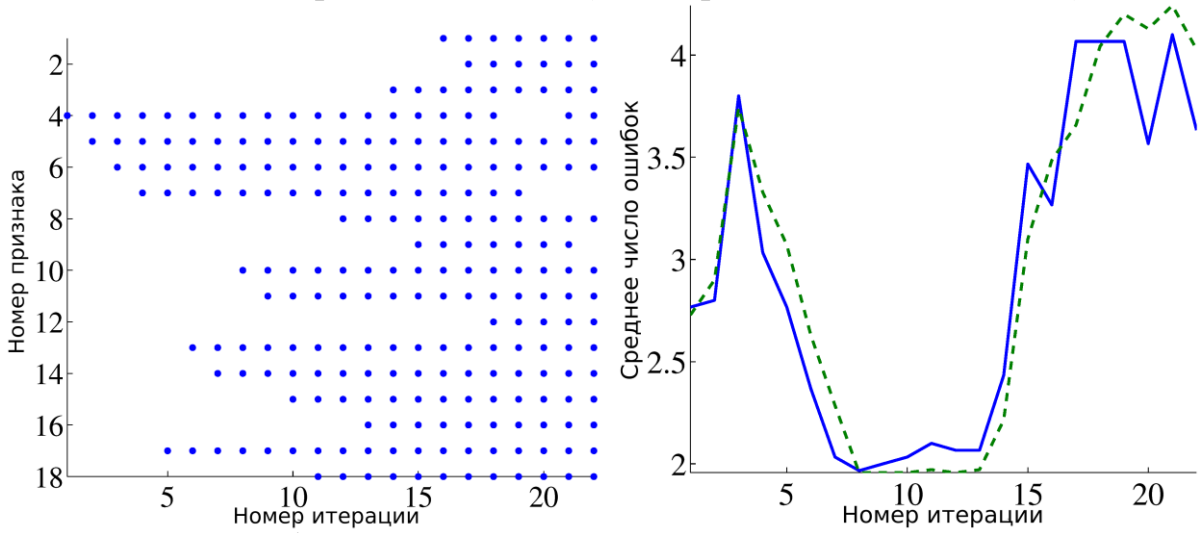
Алгоритм выбора признаков протестирован на выборке из 100 библиографических записей. Задано число используемых типов полей структуры BibTeX для данной выборки, $s = 11$: автор, название работы, название источника, номера страниц, номер выпуска, номер тома, год, город, издательство, редакторы, ссылка на работу в интернете. Максимальное число текстовых сегментов $m = 9$. Для каждого сегмента генерировался столбец из $p = 18$ признаков: длина сегмента, порядковый номер, число различных знаков препинания (точки, запятые, тире, кавычки, скобки, двоеточия, точки с запятой), общее число сегментов объекта, число заглавных букв, число цифр, количество слов, наличие инициалов, наличие подряд идущих цифр (числа), наличие подряд идущих заглавных букв (аббревиатуры), общая длина всех сегментов в записи, общее число слов в записи.

Таким образом задана матрица $\mathbf{X}$ размера $m \times n \times p$, где $m = 100$, $n = 9$, $p = 18$. Матрица ответов $\mathbf{Y}$ задана в виде (1). Множество объектов $\{\mathbf{X}_i, \mathbf{Y}_i\}$ разбито на обучающую и контрольную выборки.

На рис. 2а показан набор признаков на каждой итерации алгоритма выбора признаков. Видно, что оптимальный набор был найден за небольшое число итераций ($t = 10\ldots15$). Показано, что в данном случае некоторые признаки оказались неинформативны и были удалены сразу же после их добавления в набор. На рис. 2б показано среднее число ошибок на обучающей выборке (сплошная линия) и контрольной выборке (штриховая линия).

На рис. 3 показано количество совпадений экпертной сегментации с

сегментацией, проведенной предложенным алгоритмом: красный цвет означает нахождение в данном квадрате максимального числа сегментов, синий – минимального. Нахождение сегмента на диагонали означает совпадение экспертной и полученной сегментации. Видно, что первые по порядку поля имеют хорошее качество сегментации. Последующие поля имеют заметно худшее качество сегметации. Разница возникает в силу того, что первые поля (автор, название, год и др.) чаще присутствуют в библиографических записях и выборка достаточно велика, чтобы получить адекватную оценку вектора параметров логистической регрессии. Последние же поля присутствовали лишь в небольшом количестве из выборки 100 записей (некоторые – меньше, чем в 10).



а) Используемые признаки б) Среднее число ошибок

Рис. 2. Зависимость включения признаков в модель и среднего числа ошибок от номера итерации

Библиографические записи были разделены на 6 заданных кластеров – типов библиографических записей: статья, книга, тезис конференции, диссертация, электронный источник, либо ни один из указанных типов. Параметры λ_i метрики ρ заданы в зависимости от частоты вхождения i-го поля в библиографические записи. Чем чаще поле присутствует в библиографической записи, тем менее оно важно:

$$\lambda_i^2 = \left(1 - \frac{\sum_{j=1}^{n} B(i,j)}{n}\right)^2, \quad 0 \le \lambda_i < 1, \quad i = 1,\ldots,s.$$

Например, поле «название» или «автор» присутствует почти во всех типах записей, а «число страниц», «название журнала» – в заметно меньшем числе.

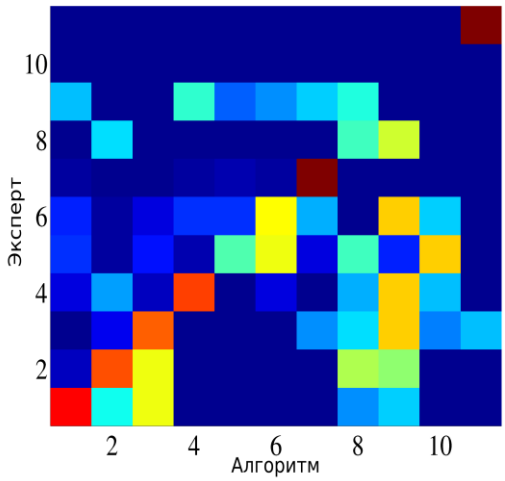
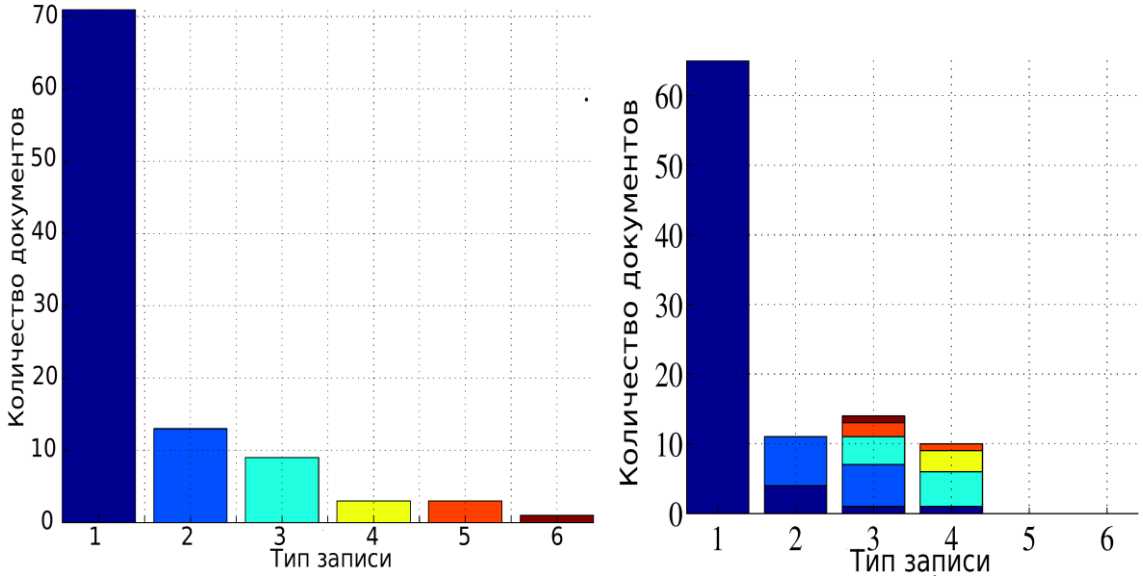


Рис. 3. Совпадение экспертной и полученной сегментации

Начальное положение центра кластера $\boldsymbol{\mu}_j$ задано теми векторами $\mathbf{b}_i = \boldsymbol{\mu}_j$, которые соответствуют записям, содержащим редко встречающиеся типы полей, и при этом находящимся на удалении друг от друга больше заданной величины $\rho_{\min}$.



а) экспертная б) полученная

Рис. 4. Кластеризации записей

Диаграмма 4а отражает количество записей, относящихся к каждому из типов BibTeX по мнению эксперта. На диаграмме 4б представлен результат кластеризации. Типы записей, представленные малым числом примеров, не определились: как видно из рис. 4б, столбцы 5 и 6 имеют нулевые значения. На тех типах записей, которые были представлены в выборке бóльшим числом примеров, предложенный метод показал результат, слабо отличающийся от экспертной кластеризации.

Заключение

В работе решена задача разметки библиографических записей. Предлагается

новая постановка задачи структурного обучения для прогнозирования структуры библиографической записи. Поставленная задача прогнозирования требует использования алгоритма выбора оптимального набора признаков. Использован алгоритм последовательного выбора признаков, исследованы его свойства. Алгоритм проиллюстрирован выборкой из неформатированных библиографических записей, для которых каждому сегменту ставилось в соответствие поле в структуре BibTeX. Для каждой записи определялся её тип в данной структуре. Представлены результаты работы предложенного метода сегментирования и кластеризации библиографических записей.

Работа выполнена при поддержке РФФИ, грант № 14-07-31326.

Список литературы

[1] *Библиографические записи в формате BibTeX.* URL: http://www.bibtex.org (дата обращения: 20.12.2012).

[2] Полежаев, В. Задачи и методы автоматического построения графа цитирований по коллекции научных документов // Труды МФТИ, 2012. Т. 4. С. 1–12.

[3] Borkar V., Deshmukh K., Saravagi S. *Automatic segmentation of text into structured records* // Proceedings of the 2001 ACM SIGMOD international conference on management of data, 2001. V. 30. N. 2. Pp. 175–186.

[4] *Cititation Parser*. URL: http://freecite.library.brown.edu/ (дата обращения: 20.12.2012).

[5] Lampert C. H. *Maximum Margin Multi-Label Structured Prediction* // Advances in Neural Information Processing Systems, 2011. V. 24. P. 289–297.

[6] Jaakkola T., Sontag D. *Learning Bayesian Network Structure using LP Relaxations* // Proceedings of the 13th International Conference on Artifcial Intelligence and Statistics, 2010. V. 9. N. 1. P. 358–365.

[7] Kwok T. Y., Yeung D. Y. *Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems* // IEEE Transactions on Neural Networks, 1997. V. 8. P. 630–645.

[8] Martins A. F. T., Dipanjan D. *Stacking Dependency Parsers* // Conference on Empirical Methods in Natural Language Processing, 2008. P. 157–166.

[9] Jaakkola T. *Scaled structured prediction.* URL: http://video.yandex.ru/users/ya-events/view/486/user-tag/научный%20семинар/ (дата обращения: 20.12.2012).

[10] Strijov V. V., Krymova E. A., Weber G. W. *Evidence optimization for consequently generated models* // Mathematical and Computer Modelling, 2013. V. 57(1-2). P. 50–56.

[11] Стрижов В. В., Крымова Е. А. Выбор моделей в линейном регрессионном анализе // Информационные технологии, 2011. Вып. 10. С. 21–26.

[12] Bishop C. M. Pattern *Recognition and Machine Learning*. LLC: Springer Science, 2006. 638 p.

[13] Адуенко А. А., Кузьмин А. А., Стрижов В. В. *Выбор признаков и оптимизация метрики при кластеризации коллекции документов* // Известия Тульского государственного университета, Естественные науки, 2012. Вып. 4, Стр. 119-131.

Bibliography

[1] Bibliograficheskie zapisi v formate BibTeX. URL: http://www.bibtex.org (reference date: 20.12.2012).

[2] Polezhaev, V. Automated citation graph building from a corpora of scientific documents // Trudy MFTI, 2012. V. 4. P. 1–12.

[3] Borkar V., Deshmukh K., Saravagi S. Automatic segmentation of text into structured records // Proceedings of the 2001 ACM SIGMOD international conference on management of data, 2001. V. 30. N. 2. P. 175–186.

[4] Cititation Parser. URL: http://freecite.library.brown.edu/ (reference date: 20.12.2012).

[5] Lampert C. H. Maximum Margin Multi-Label Structured Prediction // Advances in Neural Information Processing Systems, 2011. V. 24. P. 289–297.

[6] Jaakkola T., Sontag D. Learning Bayesian Network Structure using LP Relaxations // Proceedings of the 13th International Conference on Artifcial Intelligence and Statistics, 2010. V. 9. N. 1. P. 358–365.

[7] Kwok T. Y., Yeung D. Y. Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems // IEEE Transactions on Neural Networks, 1997. V. 8. P. 630–645.

[8] Martins A. F. T., Dipanjan D. Stacking Dependency Parsers // Conference on Empirical Methods in Natural Language Processing, 2008. P. 157–166.

[9] Jaakkola T. Scaled structured prediction. URL: http://video.yandex.ru/users/ya-events/view/486/user-tag/nauchnyj%20seminar/ (reference date: 20.12.2012).

[10] Strijov V. V., Krymova E. A., Weber G. W. Evidence optimization for consequently generated models // Mathematical and Computer Modelling, 2013. V. 57(1-2). P. 50–56.

[11] Strizhov V. V., Krymova E. A. Model Selection in Linear Regression Analysis // Informacionnye tehnologii, 2011. N. 10. P. 21–26.

[12] Bishop C. M. Pattern Recognition and Machine Learning. LLC: Springer Science, 2006. 638 p.

[13] Aduenko A. A., Kuzmin A. A., Strizhov V. V. Feature selection and metrics' optimisation when clustering documents collection // Izvestija Tul'skogo gosudarstvennogo universiteta, Estestvennye nauki, 2012. N. 4, P. 119-131.