УДК 519.767/519.237.8

КЛАССИФИКАЦИЯ ФРАГМЕНТОВ ТЕКСТОВ С ОПИСАНИЕМ ЗАВИСИМОСТЕЙ ПРАВИЛАМИ НА ИНТЕРПРЕТИРУЕМОМ ЭКСПЕРТАМИ ЯЗЫКЕ

П. А. Прокофьев

ООО «ЛАН-ПРОЕКТ»

Поступила в редакцию ..2012 г.

Аннотация. Ключевым моментом методов извлечения информации из текстов является классификация фрагментов текста. Желательно, чтобы правила извлечения были описаны на понятном экспертам языке и могли быть изменены экспертами вручную. В данной работе предлагается подход, основанный на построении дискретных процедур распознавания. Настройка процедур распознавания сопровождается автоматическим построением правил извлечения фрагментов.

Ключевые слова: извлечение информации, текст, фрагмент текста, язык правил, классификация, дискретные процедуры распознавания, выбор признаков, географическая привязка текстов.

Annotation. Fragment classification is main step in information extracting tasks for texts. It is desirable that the extraction rules are described in plain language and experts be able to change them manually. In this paper we propose an approach based on constructing discrete recognition procedures. Setting up automatic recognition procedures followed by the construction rules extraction of fragments.

Keywords: information extraction, text fragment, language of rules, classification, discrete procedure of recognition, feature selection, geo-referenced texts.

ВВЕДЕНИЕ

Задача извлечения информации из текста относится к разряду тех, для которых практически невозможно построить математическую модель в общепринятом смысле. Частным случаем задачи извлечения информации из текста является отнесение фрагментов в тексте к одному или нескольким заранее определенным классам.

В работе извлечение информации использовалось для разрешения неоднозначностей выделения географических объектов в текстах. Слова и словосочетания в текстах распределялись по классам: страна, область, район, город и прочие. Используемые в работе методы строили модели, анализ которых для экспертов достаточно сложен.

Другой подход, основанный на написании правил экспертами вручную, показывает, что правила быстро разрастаются и становятся плохо понятными самим экспертам.

В работах [1,4] используются методы автоматического построения логических правил на этапе настройки классификатора. Правила позволяют отнести целый текст или его части к определенной тематике. Язык описания правил достаточно узок. Все это делает проблематичным использование этих методов в задаче извлечения информации в текстах.

В работе предлагается метод формального описание правил извлечения информации. По правилам строятся наборы фрагментов текстов. Подобный подход разрабатывается в данной работе, но в большей степени уделяется внимание автоматизации процесса построения правил.

В рамках данной работы предлагается формальное описание языка правил извлечения фрагментов текста и метод автоматического построения правил при настройке алгоритмов классификации, используемых при извлечении информации в тексте. Алгоритмы классификации строятся с помощью дискреционных процедур распознавания по прецедентам, описанных в работах [6, 7, 8].

Методы рассмотренные в настоящей работе оценивались при решении задачи разрешения неоднозначностей при географической привязке текстов.

МОДЕЛЬ ТЕКСТОВ, ФРАГМЕНТОВ И ПРАВИЛ

Объектами исследования настоящей работы являются тексты и фрагменты в текстах. Определим формально эти понятия.

Определение 1. *Текстами* будем называть конечные последовательности слов: $\tau = (\tau_1,\ldots,\tau_L)$, где τ_i – слова, $\tau_i \in W$, $i \in \{1,\ldots,L\}$, W – множество допустимых слов, $L = L(\tau)$ – длина текста. Текст нулевой длины обозначим Λ. Обозначим множество всех текстов $\mathbb{T}$.

Определение 2. *Фрагментом текста* τ называется совокупность трех объектов: τ, i, j, для которых выполняется условие $1 \le i \le j \le L(\tau)$. Фрагмент текста будем обозначать $\tau[i,j]$, где i *– начало фрагмента,* j *– конец фрагмента.* Обозначим $\mathbb{F}(\tau)$ – множество всех фрагментов текста τ, $\mathbb{F}$ – множество всех фрагментов текстов.

Для фрагмента $S = \tau[i,j]$ число $j-i+1$ будем называть длиной фрагмента и обозначать $|S|$.

Если фрагмент $S_1 = \tau[i_1, j_1]$ включает в себя фрагмент $S_2 = \tau[i_2, j_2]$, то есть $i_1 \le i_2 \le j_2 \le j_1$, то будем писать $S_1 \sqsupset S_2$.

Если фрагмент $S_1 = \tau[i_1, j_1]$ расположен слева от фрагмента $S_2 = \tau[i_2, j_2]$, то есть $i_1 \le j_1 < i_2 \le j_2$, то будем писать $S_1 < S_2$.

Для фрагментов $S_1 = \tau[i_1, j_1]$, $S_2 = \tau[i_2, j_2]$, таких что $S_1 < S_2$ и $n \le i_2 - j_1 \le m$, будем писать $S_1 <_{n,m} S_2$.

Для фрагментов $S_1 = \tau[i_1, j_1]$ и $S_2 = \tau[i_2, j_2]$ определим бинарную операцию $S_1 \Delta S_2 = \tau[min\{i_1, i_2\}, max\{j_1, j_2\}]$.

ПРАВИЛА ИЗВЛЕЧЕНИЯ ФРАГМЕНТОВ

Определение 3. *Правилом извлечения фрагментов* (или просто правилом) будем называть отображение q, ставящее в соответствие любому тексту τ конечный набор фрагментов этого текста:

$$q(\tau) = \left\{\tau[i_1, j_1], \ldots, \tau[i_c, j_c]\right\} \subset \mathbb{F}(\tau).$$

Рассмотрим несколько правил, которые будут использоваться в настоящей работе:

1) $q^*(\tau) = \mathbb{F}(\tau)$ возвращает все фрагменты текста;

2) $q^{[n,m]}(\tau) = \left\{f \in \mathbb{F}(\tau) \mid n \le |f| \le m\right\}$, возвращает фрагменты ограниченной длины;

3) $q^{(g,\alpha)}(\tau) = \left\{\tau[i,j] \mid g(\alpha, (\tau_i, \ldots, \tau_j)) = 1\right\}$, возвращает фрагменты, последовательность слов в которых делает истинным предикат $g : A \times \mathbb{T} \to \{0,1\}$, зависящий от параметра, значение которого фиксируется значением $\alpha \in A$.

Чтобы задание правил сделать конструктивным, рассмотрим ряд операций, позволяющих получать новые правила из существующих:

1) пересечение:

$$(q_1 \Diamond q_2)(\tau) = q_1(\tau) \cap q_2(\tau);$$

2) объединение (операция «ИЛИ»):

$$(q_1 \nabla q_2)(\tau) = q_1(\tau) \cup q_2(\tau);$$

3) операция «И»

$$(q_1 \Delta q_2)(\tau) = \left\{S_1 \Delta S_2 \mid S_1 \in g_1(\tau), S_2 \in g_2(\tau)\right\};$$

4) последовательность с ограничением на расстояние:

$$(q_1 \square_{n,m} q_2)(\tau) = \left\{S_1 \Delta S_2 \mid S_i \in q_i(\tau), S_1 <_{n,m} S_2\right\};$$

5) дополнение:

$$(\neg q)(\tau) = \mathbb{F}(\tau) \setminus q(\tau);$$

6) унарная и бинарная операции префиксного условия:

$$(\square^{\leftarrow} q)(\tau) = \left\{S \mid \exists S_1 \in q(\tau), S_1 < S\right\},$$

$$q_1 \square^{\leftarrow} q_2 = (\square^{\leftarrow} q_1) \Diamond q_2;$$

7) аналогично префиксному условию задается суффиксное:

$$(\square^{\rightarrow} q)(\tau) = \left\{S \mid \exists S_1 \in q(\tau), S < S_1\right\},$$

$$q_1 \square^{\rightarrow} q_2 = q_1 \Diamond (\square^{\rightarrow} q_2);$$

8) аналогично 4), 6) и 7) задаются условия с ограничением на расстояние: $q_1 \square^{\rightarrow}_{n,m} q_2$ и $q_1 \square^{\leftarrow}_{n,m} q_2$.

ЗАДАЧА КЛАССИФИКАЦИИ ФРАГМЕНТОВ ТЕКСТА

В настоящей работе будет рассмотрен частный случай задачи извлечения информации в тексте, когда необходимо классифицировать фрагменты в тексте по нескольким заранее определенным классам.

Пусть известно, что множество фрагментов текстов $\mathbb{F}$ представимо в виде объединения непересекающихся классов $\mathbb{K}_1, \ldots, \mathbb{K}_m$. Имеется конечный набор фрагментов $X = \{S_1, S_2, \ldots, S_t\} \subset \mathbb{F}$, для которых известна их принадлежность к классам. Введем обозначения $K_i = \mathbb{K}_i \cap X$, $K_i = X \setminus K_i$.

ПРИЗНАКИ ФРАГМЕНТОВ ТЕКСТОВ

Признак фрагмента задается с помощью правила q:

$$f^{(q)}(\tau[i,j]) = \begin{cases} 1, \text{если } \tau[i,j] \in q(\tau), \\ 0, \text{ иначе.} \end{cases}$$

Вектор значений признаков, соответствующих поднабору правил $H = \{q_{i_1},\dots,q_{i_r}\} \subset R$, для фрагмента S будем называть подописанием и обозначать $(S,H) = \left(f^{(q_{i_1})}(S),\dots,f^{(q_{i_r})}(S)\right)$.

Близость подописаний фрагментов S и S' по поднабору правил H будем оценивать величиной:

$$B(S,S',H) = \begin{cases} 1, f^{(q_{i_u})}(S) = f^{(q_{i_u})}(S'), \forall u \in 1,\dots,r, \\ 0, \text{ иначе.} \end{cases}$$

ПОСТРОЕНИЕ НАБОРА ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Изначально набор правил строится из простых (базовых) правил определенного вида. В настоящей работе формируется набор базовых правил по контексту фрагмента с ограничением трех типов: на сам фрагмент, на префикс и на суффикс. Ограничения задаются предикатами. В настоящей работе использовались предикаты следующих типов:

1) $g_1(\omega,(\tau_i)) = 1$, только если в тексте τ i-е слово равно ω;

2) $g_2(\omega,(\tau_i)) = 1$, только если в тексте τ исходная морфологическая форма i-го слова равна ω;

3) $g_3(d,(\tau_i)) = 1$, только если в тексте τ у i-го слова есть дескриптор d, например, \$FirstUp (слово с большой буквы);

4) $g_4(D,(\tau_i)) = 1$, только если в тексте τ исходная морфологическая форма i-го слова принадлежит множеству D (словарю);

Из всех построенных для контекстов фрагментов правил выбирается оптимальный в некотором смысле поднабор. В настоящей работе использовалась жадная стратегия **Add-Del**, заключающаяся в итерационном добавлении в набор, а затем удалении из набора нескольких правил так, чтобы найти набор, на котором некоторый функционал качества $\lambda(R)$ имел бы оптимальное значение. Критерий отдает предпочтение коротким наборам, хорошо разделяющим классы выборки и состоящим из правил информативных по отдельности и по парам. Описание критерия выходит за рамки этой работы.

ДИСКРЕТНЫЕ ПРОЦЕДУРЫ КЛАССИФИКАЦИИ

Зафиксируем порядок отобранных правил $R = \{q_1,\dots,q_l\}$ и рассмотрим множества векторов значений признаков $\tilde{K}_i = f^{(R)}(K_i) \setminus f^{(R)}(\overline{K}_i)$. Множество $\tilde{X} = \tilde{K}_1 \sqcup \dots \sqcup \tilde{K}_m$ будем далее называть обучающей выборкой.

В работах [6,7] описаны методы построения дискретных процедур распознавания (классификации). Приведем здесь основные понятия необходимые для изложения результатов настоящей работы. Опишем дискреционные процедуры классификации на языке логических функций.

Будем называть *элементарным классификатором* элементарную конъюнкцию $c = x_{i_1}^{\alpha_1} \cdot \dots \cdot x_{i_r}^{\alpha_r}$ от переменных $x_{i_1},\dots,x_{i_r}$. Для вектора $\vec{\beta} = (\beta_1,\dots,\beta_l)$ будем обозначать $B(c,\vec{\beta}) = \beta_{i_1}^{\alpha_1} \cdot \dots \cdot \beta_{i_r}^{\alpha_r}$. Будем говорить, что вектор $\vec{\beta}$ удовлетворяет элементарному классификатору c, если $B(c,\vec{\beta}) = 1$.

Дискретная процедура классификации A состоит из множеств элементарных классификаторов $C^A(\tilde{K}_1),\dots,C^A(\tilde{K}_m)$ для каждого класса, а также весов элементарных классификаторов $\gamma(c)$, для всех $c \in C^A(\tilde{K}_i)$.

В работе описаны процедуры состоящие из элементарных классификаторов, обладающих определенными свойствами.

Элементарный классификатор называется *представительным набором* для класса $\tilde{K}_i$, если ему не удовлетворяет ни один вектор из $\tilde{X} \setminus \tilde{K}_i$.

Элементарный классификатор называется *антипредставительным набором* для класса $\tilde{K}_i$, если ему не удовлетворяет ни один вектор из $\tilde{K}_i$.

В настоящей работе также рассматриваются процедуры, состоящие из классификаторов обоих типов. Именно такие процедуры имеют наилучшие показатели.

При классификации вектора $\vec{\beta}$ для каждого класса $\tilde{K}_i$ вычисляются значения функции голосования:

$$\Gamma_1(\vec{\beta},\tilde{K}_i) = \sum_{c \in C^A(\tilde{K}_i)} \gamma(c) B(c,\vec{\beta}).$$

Вектор относится к классу, для которого значение функции голосования максимально.

ПОИСК ЭЛЕМЕНТАРНЫХ КЛАССИФИКАТОРОВ

В работе описан способ нахождения представительных и антипредставительных наборов,

как элементарных конъюнкций допустимых для частично определенной логической функции. Такие элементарные конъюнкции называются *импликантами* функции. Рассмотрим этот способ на примере голосования по представительным наборам.

Пусть для класса $\tilde{K}_i$ частично определена на $\tilde{X}$ функция

$$u^{(\tilde{K}_i,\tilde{X})}(\vec{\beta}) = \begin{cases} 1, \vec{\beta} \in \tilde{K}_i \\ 0, \vec{\beta} \in \tilde{X} \setminus \tilde{K}_i. \end{cases} \quad (1)$$

Теоретически импликанты можно получить при нахождении сокращенной ДНФ доопределенной функции. Сокращенная ДНФ находится методами, приведенными в книге .

Однако в настоящей работе набор признаков превышает 50, объем обучающих классов превышает 10^3, а при поиске получается более 10^7 конъюнкций. В работе предлагается искать только элементарные классификаторы, обладающие определенными свойствами, зависящими от параметров алгоритма, однако нет рекомендайций по выбору этих параметров. В настоящей работе использовался предложенный автором адаптивный алгоритм, параметры которого изменяются в процессе вычислений по критерию допустимой сложности вычислений.

ВЫБОР ВЕСОВ ЭЛЕМЕНТАРНЫХ КЛАССИФИКАТОРОВ

В работе предлагается разбивать обучающую выборку на базовую и контрольную. Наборы элементарных классификаторов строятся по базовой выборке, а веса вычисляются исходя из классификации контрольной выборки.

Другой подход вычисления весов может быть основан на других обучаемых методах классификации, например, **MaxEnt**. В настоящей работе построенные по базовой выборке элементарные классификаторы для всех классов объединяются в один набор и рассматриваются как признаки, далее обучение и классификация осуществляется методом MaxEnt.

Таблица 1

Примеры отобранных автоматически «базовых» правил

Правило (q)	$\mu(q)$
@ОБЛАСТЬ	0,98
#1 :1? РАЙОН	0,36
#1 :1? КРАЙ	0,82
ЕДИНАЯ РОССИЯ	0,05

ЭКСПЕРИМЕНТЫ

Метод, предложенный в работе географической привязки текстов, заключался в классификации фрагментов, которые были найдены в географическом справочнике. Слова классифицировались по типам географических объектов, чтобы снять неоднозначность, связанную с неуникальностью имен (объекты разных типов имеют одинаковые названия).

Предложенный в настоящей работе подход был протестирован на этой задаче. Размеченный корпус состоит из 846 текстов, содержащих 372367 слов (в то числе знаков пунктуации). В текстах 5431 фрагментов было отнесено к одному из 8 типов географических объектов. При поиске в географическом справочнике было найдено 6408 «подозрительных» фрагментов, то есть 977 фрагментов были отнесены к классу «другие».

Выбор базовых правил осуществлялся по всей обучающей выборке. Было отобрано 56 правил.

Среди автоматически отобранных правил встречались вполне ожидаемые, например, приведенные в таблице 1. Правила в таблице написаны на специальном языке так, чтобы эксперты, знающие синтаксис языка, могли понять и исправить при необходимости правила. Строка «@ОБЛАСТЬ» соответствует правилу $q^{(g_4, D_{\text{обл.}})}$, где $D_{>1;.}$ – словарь названий областей России. Строка «#1 :1? РАЙОН» соответствует правилу $q^{[1]}\square^{\rightarrow}_{1,1}q^{(g_2,\text{район})}$.

Таблица 2

Оценка качества методов

Метод	P	R	F	e
ME	42,5%	37,6%	35,2%	17,4%
Γ_1	76,2%	65,8%	63,3%	15,6%
Γ_2	38,3%	24,8%	21,7%	27,2%
Γ_1 + *ME*	76,1%	59,8%	54,9%	22,2%
Γ_2 + *ME*	30,0%	21,3%	19,9%	25,2%
$\Gamma_1 + \Gamma_2$	77,1%	72,6%	69,6%	11,2%

Таблица 3

Оценка качества для хорошего класса «область»

Метод	P	R	F	e
ME	61,3%	86,9%	71,9%	7,7%
Γ_1	94,3%	98,0%	96,2%	0,9%
Γ_2	0,0%	0,0%	0,0%	12,0%
Γ_1 + *ME*	95,4%	98,8%	97,1%	0,7%
$\Gamma_1 + \Gamma_2$	94,3%	98,0%	96,2%	0,3%

В настоящей работе были оценены показатели качества для следующих методов:

1) метод классификации MaxEnt (*ME*);

2) по представительным наборам (Γ_1);

3) по антипредставительным наборам (Γ_2);

4) MaxEnt с использование представительных наборов ($\Gamma_1 + ME$);

5) MaxEnt с использование антипредставительных наборов ($\Gamma_2 + ME$);

6) голосование по представительным и антипредставительным ($\Gamma_1 + \Gamma_2$);

Для каждого метода проводилось по 3 итерации. При обучении на каждой итерации выборка случайно делилась на тестовую (30%) и обучающую (70%). При оценке методов Γ_1 и Γ_2 обучающая выборка разбивалась на базовую (50%) и проверочную (50%) для вычисления весов по формуле (**).

При оценке каждого метода вычислялись показатели точности (P), полноты (R), F -меры (F), а также процент ошибки (e) для каждого класса, а также макро-усреднения этих параметров. Макро-усреднения оценок приведены в таблице 2. Худшие результаты в среднем показал метод голосования по антипредставительным наборам. Для больших классов были построены длинные антипредставительные наборы, которые на тестовой выборке голосовали очень редко, и были получены показатели $P = 0$ и $R = 0$.

Лучшие результаты показал метод голосования по представительным и антипредставительным наборам, в котором складываются голоса классификатором из процедур Γ_1 и Γ_2.

Примером сложного построенного правила является «@ОБЛАСТЬ :1! РАЙОН», где оператор «:1!» означает отрицание суффиксного правила. Анализ элементарных классификаторов может быть автоматизирован для построения более сложных правил.

ВЫВОДЫ

В работе дается формальное описание конструкции правил извлечения фрагментов текста. Показывается принципиальная возможность представления элементарных классификаторов в виде правил. Дальнейшие исследования будут направлены на расширение и строгое формальное описание языка правил, разработку методов построения правил по элементарным классификаторам, исследование алгоритмов автоматического построения правил.

СПИСОК ЛИТЕРАТУРЫ

1. *Junker M., Abecker A.* Learning Complex Patterns for Document Categorization // AAAI-98/ ICML Workshop on Learning for Text Categorization. Madison, Wisconsin, USA, 1998.

2. *Reiss F., Raghavan S., Krishnamurthy R., Zhu H., Vaithyanathan S.* An Algebraic Approach to Rule-Based Information Extraction // ICDE. Cancun, Mexico, 2008.

3. *Ratnaparkhi A.* Maximum Entropy Models for Natural Language Ambiguity resolution. PHD thesis, Univ. of Pennsylvania, 1998.

4. *Агеев М.С.* Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов. – Диссертация на соискание ученой степени к.ф.-м.н. – М.: МГУ, 2004.

5. *Дюкова Е.В., Песков Н.В.* Построение распознающих процедур на базе элементарных классификаторов // Математические вопросы кибернетики / Под ред. О.Б. Лупанов. – М.: Физматлит, 2005. – Т. 14.

6. *Дюкова Е.В.* Дискретные (логические) процедуры распознавания: принципы конструирования, сложность реализации и основные модели. Учебное пособие для студентов математических факультетов педвузов. – М.: Изд-во «Прометей», 2003. – 29 с.

7. *Песков Н.В.* Поиск информативных фрагментов описаний объектов в задачах распознавания. – Диссертация на соискание ученой степени к.ф.-м.н., М.: ВЦ РАН. – 2004.

8. *Прокофьев П.А.* Использование методов извлечения информации при географической привязке текстов на русском языке // Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции (RCDL). Петрозаводск, 2009.

9. Дискретная математика и математические вопросы кибернетики / Под ред. С.Б. Яблонского, О.Б. Лупанова, М.: «Наука», 1974. 312 с.

Прокофьев П.А. – научный консультант ООО «ЛАН-Проект». Тел. +7(926)2176188, e-mail: p_prok@mail.ru.

Prokofyev P.A. – research adviser Company "LAN-project." Tel. +7(926)2176188, e-mail: p_prok@mail.ru.