# Классификация и выделение фрагментов в текстах на основе логических правил

© В.Г. Васильев

ООО «ЛАН-ПРОЕКТ»
vvg_2000@mail.ru

**Аннотация**

В работе рассматривается подход к классификации и выделению значимых фрагментов текстов, основанный на использовании математической модели, в которой логические правила определяются в виде операции над множествами фрагментов. Проводится подробный анализ алгебраических свойств различных вариантов определения правил. Особое внимание уделяется исследованию свойств операций, в которых задаются ограничения на расстояния между фрагментами. Разрабатываются и обосновываются эффективные алгоритмы реализации правил.

## 1. Введение

В настоящее время во многих крупных организациях для обеспечения аналитической работы создаются хранилища корпоративных и открытых документов (электронные библиотеки). Важным элементом, позволяющим автоматизировать обработку и анализ документов, в таких хранилищах, являются средства автоматической классификации.

Для их построения используются два основных подхода: обучение на примерах и задание правил на специальном языке. При этом обычно рассматривается задача классификации текстов целиком без выделения в них значимых фрагментов [4,6,7]. Во многих практических случаях это является существенным ограничением.

В работе [8] была рассмотрена комплексная технология классификации и выделения фрагментов в текстах, реализованная в системе СКАТ, а в работе [9] описан используемый в ней подход к выделению фрагментов, основанный на обучении на примерах документов. Данный подход хорошо работает при наличии достаточного количества примеров документов, а также при относительно простой структуре рубрик классификатора. Однако на практике при обучении на примерах возникает ряд трудностей, связанных с отсутствием достаточного количества примеров документов и наличием различных неточностей и ошибок [2,7,8]. В этой ситуации наряду с обучением на примерах требуется явное задание правил классификации на специальном языке.

В ряде систем, например, Oracle, Артефакт, ABBYY, Reuters Construe, реализованы достаточно развитые языки задания правил классификации, однако в результате их работы в текстах выделяются отдельные слова и словосочетания, а алгебраические свойства реализованных операций формально не описываются и не исследуются [4,6]. В работах [1,3] проводится подробный анализ алгебраических свойств стандартных логических операций (И, ИЛИ, НЕ) при выделении фрагментов в текстах. Однако особенности и свойства операций, в которых задаются ограничения на расстояния между элементами текста, остаются не исследованными.

В настоящей работе рассматриваются модели представления текстов и правил, которые реализованы в системе СКАТ и специально ориентированы на выделение и задание операций над фрагментами.

За счет построения формальной математической модели удается провести анализ алгебраических свойств как стандартных логических операций, так и операций, в которых задаются ограничения на расстояния, размер и взаимное расположение фрагментов в текстах; определить условия при которых операции будут обладать ассоциативности и дистрибутивности, а также при которых возможна их эффективная реализация.

## 2. Модели представления текстов и задания правил

### 2.1 Основные определения и понятия

Для построения модели текстов и правил классификации воспользуемся подходом, основанном на представлении текста в виде множества фрагментов [1].

Пусть имеется текст $D$, являющийся последовательностью элементов (слов, цифр, знаков препинания), т.е. $D = (d_1, \ldots, d_n)$, где $d_i \in T$ – отдельный элемент текста, $T = \{t_1, \ldots, t_m\}$ – множество всех допустимых элементов, $n$ – длина текста, $m$ – число различных допустимых элементов текстов.

**Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011**

**Определение.** Множество $\mathbb{F} = \{(p,q)|1 \le p \le q \le n\}$ будем называть множеством всех фрагментов текста длины $n$. Фрагментами текста будем называть отдельные элементы данного множества $f = (f_l, f_r) \in \mathbb{F}$, которые задают левую $f_l$ и правую $f_r$ границы фрагмента (номер начального и конечного элемента текста). ■

**Определение.** Пусть $f = (f_l, f_r) \in \mathbb{F}$ и $g = (g_l, g_r) \in \mathbb{F}$ тогда:

$|f| \equiv f_r - f_l + 1$ – длина фрагмента;

$g \sqsupset f$, если $g_l \le f_l \le f_r \le g_r$ и $f \ne g$ – отношение включения;

$f \ll g$, если $f_l < g_l$ или $f_l = g_l \wedge f_r < g_r$ – отношение упорядочения фрагментов. ■

**Определение.** Результатом выполнения произвольного правила $Q$ для текста $D$ является множество $F_Q \subset \mathbb{F}$, содержащее все фрагменты удовлетворяющие правилу $Q$. При этом, если $F_Q \ne \emptyset$, то будем говорить, что текст $D$ удовлетворяет правилу $Q$. ■

При таком подходе результирующее множество фрагментов $F_Q$ является избыточным и содержит большое количество вложенных друг в друга фрагментов. Действительно, если фрагмент $f \in F_Q$, то все фрагменты $g \in \mathbb{F}$, которые содержат $f$, также принадлежат $F_Q$.

По этой причине для описания результатов выполнения правил можно перейти к рассмотрению редуцированных множеств фрагментов, которые не содержат вложенных друг в друга фрагментов.

**Определение.** Множество фрагментов $F \subset \mathbb{F}$ редуцированное, если не существует $f, g \in F$ таких, что $f \sqsupset g$ или $g \sqsupset f$ .

**Определение.** Пусть имеется произвольное множество фрагментов $F \subset \mathbb{F}$, тогда $R(F) = \{f|f \in F, \nexists g \in F, f \sqsupset g\}$ – обозначает редуцированное множество фрагментов, полученное на основе $F$, а $R$ - операцию редуцирования. ■

**Замечание.** Если $F$ редуцированное, то $F = R(F)$.

Можно показать, что редуцированные множества обладают следующими свойствами [1]:

1. Элементы редуцированного множества фрагментов $A$ одинаково упорядочиваются как по началам, так и по концам фрагментов.

2. Произвольное редуцированное множество фрагментов $A$ для документа длины $n$ содержит не более $n$ элементов, т.е. $|A| \le n$.

3. Если $A$ произвольное множество фрагментов и $f \in A$, то $\exists f' \in R(A)$, такой, что $f \sqsupset f'$.

4. Если $A \subset B$, где B – редуцированное множество фрагментов, то $A$ – редуцированное множество фрагментов.

### 2.2 Модель правил для редуцированных множеств фрагментов

Правила классификации, основанные на использовании редуцированных множеств, можно определить следующим образом.

1. Элементарным правилом будем называть правило $Q = t$, $t \in T$, результат которого $F_Q = \{f_1, \dots, f_l\}$ – произвольное редуцированное множество фрагментов, элементы которого выделяются с помощью одной операции.

2. Сложным правилом будем называть такое правило $Q$, которое получено путем выполнения операций над другими правилами $Q_1, \dots, Q_k$.

3. Единичным правилом $I$ будем называть такое правило $Q$, результат которого $F_I = \{(1,1), \dots, (l,l)\}$. Пустым правилом $E$ будем называть такое правило $Q$, результат которого $F_E = \emptyset$. Заметим, что $F_I$ и $F_E$ – редуцированные множества.

В системе СКАТ элементарными правилами являются такие, которые выделяют отдельные слова в тексте, предложения, строки, разделы документа. Например, правило *$FirstUp* – выделяет все слова в тексте с большой буквы, правило *Липецкая* – все слова, являющиеся словоформами слова «липецкая», правило *‘обл*’* – все слова начинающиеся на «обл»; *$Sentence* – все предложения в документе, *#section* – раздел документа с определенным именем (например, заголовок).

**Определение.** Пусть имеется два фрагмента $f = (f_l, f_r) \in \mathbb{F}$ и $g = (g_l, g_r) \in \mathbb{F}$, тогда расстояние между ними

$$d(f,g) \equiv \begin{cases} g_l - f_r, & f < g, \\ f_l - g_r, & g < f, \\ g_l - f_r & g = f. \end{cases}$$

Приведем теперь определения возможных операций для построения сложного правила $Q$ на основе правил $Q_1, \dots, Q_k$.

1. $Q = Q_1 \nabla Q_2$ – бинарная операция ИЛИ, $F_Q \equiv R\left(F_{Q_1} \nabla^* F_{Q_2}\right)$, $F_{Q_1} \nabla^* F_{Q_2} = \{f \in \mathbb{F} | \exists f_1 \in F_{Q_1}, f \sqsupset f_1 \text{ или } \exists f_2 \in F_{Q_2}, f \sqsupset f_2\}$.

Например, правило *«Липецкие новости»* | *Липецкстрой* выделяет фрагменты, равные соответствующим выражениям.

2. $Q = Q_1 \Delta Q_2$ – бинарная операция И, $F_Q \equiv R\left(F_{Q_1} \Delta^* F_{Q_2}\right)$, $F_{Q_1} \Delta^* F_{Q_2} = \{f \in \mathbb{F} | \exists f_1 \in F_{Q_1}$ и $\exists f_2 \in F_{Q_2}$, т. что $f \sqsupset f_1$ и $f \sqsupset f_2\}$.

Например, правило *(МВД полиция «Министерство внутренних дел») & (коррупция взятка)* – выделяет фрагменты, которые содержат упоминание о МВД и коррупции.

3. $Q = Q_1 \Delta_{n_1} Q_2$ – бинарная операция И с ограничением на расстояние между фрагментами, $F_Q \equiv R\left(F_{Q_1} \Delta^*_{n_1} F_{Q_2}\right)$, $F_{Q_1} \Delta^*_{n_1} F_{Q_2} = \{f \in \mathbb{F} | \exists f_1 \in F_{Q_1}$ и $\exists f_2 \in {}_{-Q_2}$, т. что $f \sqsupset f_1, f \sqsupset f_2$ и $d(f_1, f_2) \le n_1\}$.

Например, правило *(МВД полиция «Министерство внутренних дел») &5w (коррупция взятка)* – выделяет фрагменты, которые содержат упоминание о МВД и коррупции, расстояние между которыми не более 5 слов.

4. $Q = Q_1 \square Q_2$ – бинарная операция нахождения последовательности, $F_Q \equiv R\left(F_{Q_1} \square^* F_{Q_2}\right)$, $F_{Q_1} \square^* F_{Q_2} =$

$\{f \in \mathbb{F} | \exists f_1 \in F_{Q_1}$ и $\exists f_2 \in F_{Q_2}$, т. что $f_1 < f_2, d(f_1, f_2) > 0, f \sqsupset f_1$ и $f \sqsupset f_2\}$.

Например, правило *правительство : область* – выделяет фрагменты которые начинаются на «правительство», а заканчиваются на слово «область».

5. $Q = Q_1 \square_{n_1,n_2} Q_2$ – бинарная операция последовательности с ограничением на расстояние между фрагментами, $F_Q \equiv R\left(F_{Q_1} \square^*_{n_1,n_2} F_{Q_2}\right)$, $F_{Q_1} \square^*_{n_1,n_2} F_{Q_2} = \{f \in \mathbb{F} | \exists f_1 \in F_{Q_1}$ и $\exists f_2 \in F_{Q_2}$, т. что $f_1 < f_2, d(f_1, f_2) > 0, f \sqsupset f_1, f \sqsupset f_2$ и $n_1 \leq d(f_1, f_2) \leq n_2\}$.

Например, правило (*правительство руководство) :2w (Россия «Российская Федерация» РФ)* – выделяет фрагменты, в которых правительство и название России находятся в заданном порядке и на расстоянии не более 2 слов.

6. $Q = \bowtie (Q_1, \dots, Q_k)$ – множественная операция последовательности соседних элементов (осуществляет отбор смежных фрагментов), $F_Q \equiv R\left(\bowtie^* \left(F_{Q_1}, \dots, F_{Q_k}\right)\right)$, $\bowtie^* \left(F_{Q_1}, \dots, F_{Q_k}\right) = \{f \in \mathbb{F} | \exists f_i \in F_{Q_i}, i = 1, \dots k$, т. что $f_i < f_{i+1}, d(f_i, f_{i+1}) = 1$ для $i = 1, \dots, k-1$ и $f \sqsupset f_i$ для $i = 1, \dots, k\}$.

Например, правило *«(начальник руководитель директор) («главное управление» управление организация отдел) (МВД МЧС МинФин)»* - выделяет словосочетания соответствующие руководителям различных ведомств.

7. $Q = Q_1 \lozenge Q_2$ – бинарная операция нахождения пересечения фрагментов, $F_Q \equiv \{f \in \mathbb{F} | f \in F_{Q_1} \wedge f \in F_{Q_2}\}$.

Например, правило *[великая $FirstUp]* – выделяет слова «великая», которые написаны с большой буквы.

8. $Q = Q_1 \lhd_{n_1,n_2}$ – унарная операция ограничения длины фрагмента, $F_Q \equiv \{f \in F_{Q_1} | n_1 \leq |f| \leq n_2\}$.

Например, правило *(Нижегородская & Владимирская) #IN #INTERVAL(2w/3w)* – выделяет фрагменты, содержащие заданные слова длиной от 2 до 3 слов.

9. $Q = Q_1 \lhd Q_2$ - бинарная операция проверки включения фрагмента, $F_Q \equiv \{f \in F_1 | \exists f_2 \in F_{Q_2}$, т. что $f_2 \sqsupset f\}$.

Например, правило *правительство #IN #Section(Title)* – проверяет наличие слова правительство в заголовке;

10. $Q = Q_1 \rhd Q_2$ - бинарная операция проверки содержания фрагмента, $F_Q \equiv \{f \in F_1 | \exists f_2 \in F_{Q_2}$, т. что $f \sqsupset f_2\}$.

Например, правило *$Sentence #contains (Нижегородская Владимирская)* – выделяет предложения, которые содержат слова «Нижегородская» и «Владимирская».

**Замечание.** Необходимо отметить, что в данном случае не рассматривается унарная операция НЕ. Это связано с тем, что в результате ее применения получаются вырожденные множества фрагментов. Например, для произвольного множества из одного фрагмента длины больше 1 результатом операции НЕ будет множество $F_I$.

Для возможности построения правил включающих отрицания и условные операторы (наличие выражения проверяется, но оно не включается в итоговый фрагмент) можно использовать специальные варианты бинарных правил $\nabla$, $\Delta$, $\square$, $\bowtie$, $\lozenge$, $\lhd$, $\rhd$, $\Delta_{n_1}$, $\square_{n_1,n_2}$, в которых один из операндов считается отрицательным или условным.

Например, $\square^{\nrightarrow}_{n_1,n_2}$ - операция нахождения последовательности, в которой второй операнд берется с отрицанием, $\square^{\nleftarrow}_{n_1,n_2}$ - операция, в которой первый операнд берется с отрицанием, $\square^{\leftarrow}_{n_1,n_2}$ - операция, в которой первый операнд является условным. Определение $\square^{\nrightarrow}_{n_1,n_2}$ имеет следующий вид $Q = Q_1 \square^{\nrightarrow}_{n_1,n_2} Q_2$, где $F_Q \equiv \{f \in F_{Q_1} | \exists! f_2 \in F_{Q_2}$, т. что $f < f_2, 0 < n_1 \leq d(f, f_2) \leq n_2\}$.

Редукцию результирующего множества при этом проводить не нужно, так как оно является подмножеством редуцированного множества фрагментов.

**Замечание.** По аналогии с приведенными бинарными операциями можно определить также множественные операции $\nabla(Q_1, \dots, Q_k)$, $\square(Q_1, \dots, Q_k)$, $\Delta(Q_1, \dots, Q_k)$. Свойства данных операций и особенности реализации аналогичны соответствующим бинарным операциям и по этой причине для экономии места в настоящей работе рассматриваться не будут.

**Замечание.** При определении трех последних операций не используется оператор редуцирования множества фрагментов, так как в данных случаях результирующее множество фрагментов $F_Q$ является подмножеством редуцированного множества фрагментов.

**Замечание.** Для операций с ограничениями на расстояние между фрагментами $Q_1 \Delta_{n_1} Q_2$, $Q_1 \square_{n_1,n_2} Q_2$ существенно то, что $F_{Q_1}$ и $F_{Q_2}$ являются редуцированными. В противном случае можно показать, что $Q_1 \Delta_{n_1} Q_2 = Q_1 \Delta Q_2$ и $Q_1 \square_{n_1,n_2} Q_2 = Q_1 \square Q_2$, т.е. оценка расстояния между фрагментами не имеет смысла.

### 2.3 Анализ алгебраических свойств операций для редуцированных множеств фрагментов

Рассмотрим теперь основные алгебраические свойства введенных операций для построения правил классификации.

В работе [1] показывается, что операции $\Delta$ и $\nabla$ обладают свойствами коммутативности, ассоциативности, дистрибутивности, а также имеют нейтральные и нулевые элементы равные $E$ и $I$.

Свойства бинарной операции $\square$ описываются следующим утверждением.

**Утверждение 1.** Для произвольных правил $Q_1$, $Q_2$, $Q_3$ выполняются следующие свойства:

1. ассоциативность:
$(Q_1 \square Q_2) \square Q_3 = Q_1 \square (Q_2 \square Q_3)$ ,
2. дистрибутивность относительно $\Delta$ и $\nabla$:
$(Q_1 \Delta Q_2) \square Q_3 = (Q_1 \square Q_3) \Delta (Q_2 \square Q_3)$,
$(Q_1 \square Q_2) \Delta Q_3 = (Q_1 \Delta Q_3) \square (Q_2 \Delta Q_3)$,
$(Q_1 \nabla Q_2) \square Q_3 = (Q_1 \square Q_3) \nabla (Q_2 \square Q_3)$,
$(Q_1 \square Q_2) \nabla Q_3 = (Q_1 \nabla Q_3) \square (Q_2 \nabla Q_3)$. ■

Из определения $\square(Q_1, \dots, Q_k)$ и приведенных свойств бинарной операции $Q_1 \square Q_2$ следует справедливость следующего соотношения:

$$\square(Q_1, \dots, Q_k) = \big((Q_1 \square Q_2) \dots \square Q_{k-1}\big) Q_k.$$

**Замечание.** Требование $d(f_1, f_2) > 0$ в определении операции $Q_1 \square Q_2$ является существенным для наличия свойства ассоциативности и дистрибутивности данной операции. В противном случае, если, например, $F_{Q_1} = \{(1,4)\}$, $F_{Q_2} = \{(4,5)\}$, $F_{Q_3} = \{(3,4)\}$, то $(Q_1 \square Q_2) \square Q_3 \neq Q_1 \square (Q_2 \square Q_3)$.

Рассмотрим теперь свойства бинарных операций $\square_{n_1,n_2}$ и $\Delta_{n_1}$, в которых задаются ограничения на расстояния между фрагментами.

**Утверждение 2.** Операция $\Delta_{n_1}$ имеет следующие свойства:

1. Для произвольных правил $Q_1$ и $Q_2$ справедливо свойство коммутативности, т.е. $Q_1 \Delta_{n_1} Q_2 = Q_2 \Delta_{n_1} Q_1$.

2. Свойство ассоциативности в общем случае не выполняется.

3. Свойство дистрибутивности операции $\Delta_{n_1,n_2}$ относительно $\nabla$ не выполняется. ■

**Утверждение 3.** Для операции $\square_{n_1,n_2}$ справедливы следующие свойства:

1. Свойство ассоциативности в общем случае не выполняется.

2. Если $n_1 = n_2 = n$ и $m_1 = m_2 = m$, то справедливо соотношение $\big(Q_1 \square_{n,n} Q_2\big) \square_{m,m} Q_3 = Q_1 \square_{n,n} (Q_2 \square_{m,m} Q_3)$, т.е. в данном случае свойство ассоциативности выполняется.

3. Свойство дистрибутивности операции $\square_{n_1,n_2}$ относительно $\nabla$ не выполняется.

**Следствие.** Из утверждения 3 следует, что свойством ассоциативности обладает операция $\square_{1,1}$, которая выделяет соседние фрагменты. Отсюда получаем, что

$$\bowtie (Q_1, \dots, Q_k) \equiv Q_1 \square_{1,1} Q_2 \square_{1,1} \dots \square_{1,1} Q_k.$$

Свойства операций, накладывающих ограничение на размер и включение фрагментов друг в друга описываются следующим утверждением.

**Утверждение 4.** Для произвольных правил $Q_1, Q_2, Q_3$ и $n_1, n_2, n_3, n_4 \in \mathbb{N}$ справедливы следующие свойства:

$$Q_1 \lhd_{n_1,n_2} \lhd_{n_3,n_4} = Q_1 \lhd_{n_3,n_4} \lhd_{n_1,n_2} = Q_1 \lhd_{\max(n_1,n_3),\min(n_2,n_4)},$$
$$(Q_1 \nabla Q_2) \lhd_{n_1,n_2} = \big(Q_1 \lhd_{n_1,n_2}\big) \nabla \big(Q_2 \lhd_{n_1,n_2}\big),$$
$$(Q_1 \nabla Q_2) \lhd Q_3 = (Q_1 \lhd Q_3) \nabla (Q_2 \lhd Q_3),$$
$$Q_1 \lhd (Q_3 \nabla Q_2) = (Q_1 \lhd Q_2) \nabla (Q_1 \lhd Q_3),$$
$$(Q_1 \lhd Q_2) \lhd Q_3 = (Q_1 \lhd Q_3) \lhd Q_2$$

Для правил $Q_1, Q_2$ и правила $Q_3$ такого, что $\big|F_{Q_3}\big| = 1$ справедливы следующие свойства:

$$(Q_1 \Delta Q_2) \lhd Q_3 = (Q_1 \lhd Q_3) \Delta (Q_2 \lhd Q_3),$$
$$(Q_1 \square Q_2) \lhd Q_3 = (Q_1 \lhd Q_3) \square (Q_2 \lhd Q_3),$$
$$\big(Q_1 \Delta_{n_1,n_2} Q_2\big) \lhd Q_3 = (Q_1 \lhd {}_3) \Delta_{n_1,n_2} (Q_2 \lhd Q_3),$$
$$\big(Q_1 \square_{n_1,n_2} Q_2\big) \lhd Q_3 = (Q_1 \lhd Q_3) \square_{n_1,n_2} (Q_2 \lhd Q_3).$$

■

**Утверждение 5.** Для произвольных правил $Q_1, Q_2, Q_3$ справедливы следующие свойства:

1. коммутативность:
$Q_1 \natural Q_2 = Q_2 \natural Q_1$;
2. ассоциативность:
$Q_1 \natural (Q_2 \natural Q_3) = (Q_1 \natural Q_2) \natural Q_3$. ■

Таким образом, для операции без ограничений на расстояния и размер фрагментов выполняются свойства коммутативности, ассоциативности и дистрибутивности. Однако, для операций с ограничениями на расстояния между фрагментами данные свойства в общем случае не выполняются (за исключением отдельных случаев).

### 2.4. Модифицированная модель правил

Предварительный анализ показал, что нарушение свойств ассоциативности и дистрибутивности операции $\Delta_{n_1}$ и $\square_{n_1,n_2}$ связано с потерей части информации о расположении фрагментов при редукции результатов дочерних операций. Рассмотрим модифицированную модель, в которой результатом правил являются частично редуцированные множества фрагментов.

**Определение.** Пусть $Q_1, \dots, Q_k$ произвольные правила и $Q$ сложное правило на их основе, тогда модифицированные операции определяются следующим образом.

1. Бинарная операция ИЛИ
$Q = Q_1 \nabla^+ Q_2$, $F_Q \equiv \{f \in F_{Q_1} \vee f \in F_{Q_2}\}$

2. Бинарная операция И
$Q = Q_1 \Delta^+ Q_2$, $F_Q \equiv \{(f_l, f_h) \in F | f_l = \min(f_{1l}, f_{2l})$, $f_h = \max(f_{1h}, f_{2h})$, где $f_1 \in F_{Q_1}$, $f_2 \in F_{Q_2}$ }.

3. Бинарная операция И с ограничением на расстояние между фрагментами
$Q = Q_1 \Delta^+_{n_1} Q_2$,
$F_Q \equiv \{(f_l, f_h) \in F | f_l = \min(f_{1l}, f_{2l})$, $f_h = \max(f_{1h}, f_{2h})$, где $f_1 \in F_{Q_1}$, $f_2 \in F_{Q_2}, d(f_1, f_2) \leq n_1$ }.

5. Бинарная операция нахождения последовательности
$Q = Q_1 \square^+ Q_2$, $F_Q \equiv \{(f_l, f_h) \in F | f_l = f_{1l}, f_h = -_{2h},\ f_1 \in F_{Q_1}, f_2 \in F_{Q_2},\ f_1 < f_2,\ d(f_1, f_2) > 0\}$.

6. Бинарная операция нахождения последовательности с ограничением на расстояние между фрагментами
$Q = Q_1 \square^+_{n_1,n_2} Q_2$, $F_Q \equiv \{(f_l, f_h) \in F | f_l = f_{1l}, f_h = f_{2h},\ f_1 \in F_{Q_1}, f_2 \in F_{Q_2},\ f_1 < f_2,\ 0 < n_1 \leq d(\mathrm{c}_1, f_2) \leq n_2\}$

**Замечание.** Для остальных операций, приведенных в разделе 2.2, определение остается без изменений, так как в них не выполняется редукция результирующих множеств.

### 2.5 Анализ алгебраических свойств операций для модифицированной модели

Пусть имеются правила $Q_1$, $Q_2$ и $Q_3$, которые были получены в результате применения только модифицированных операций, либо являются элементарными правилами. Тогда справедливы следующие утверждения.

**Утверждение 6.** Для операций $\nabla^+$, $\Delta^+$ и $\square^+$ справедливы следующие свойства.

1. Ассоциативность:

$$(Q_1\nabla^+Q_2)\nabla^+Q_3 = Q_1\nabla^+(Q_2\nabla^+Q_3),$$
$$(Q_1\Delta^+Q_2)\Delta^+Q_3 = Q_1\Delta^+(Q_2\Delta^+Q_3),$$
$$(Q_1\square^+Q_2)\square^+Q_3 = Q_1\square^+(Q_2\square^+Q_3).$$

2. Дистрибутивность.

$$(Q_1\nabla^+Q_2)\Delta^+Q_3 = (Q_1\Delta^+Q_3)\nabla^+(Q_2\Delta^+Q_3),$$
$$(Q_1\nabla^+Q_2)\square^+Q_3 = (Q_1\square^+Q_3)\nabla^+(Q_2\square^+Q_3). \blacksquare$$

**Утверждение 7.** Для операций $\Delta^+_{n_1}$ и $\square^+_{n_1,n_2}$ справедливы следующие свойства.

1. Ассоциативность:

$$\left(Q_1\square^+_{n_1,n_2}Q_2\right)\square^+_{m_1,m_2}Q_3 = Q_1\square^+_{n_1,n_2}\left(Q_2\square^+_{m_1,m_2}Q_3\right).$$

2. Дистрибутивность.

$$(Q_1\nabla^+Q_2)\Delta^+_{n_1}Q_3 = \left(Q_1\Delta^+_{n_1}Q_3\right)\nabla^+(Q_2\Delta^+_{n_1}Q_3),$$
$$(Q_1\nabla^+Q_2)\square^+_{n_1,n_2}Q_3 =$$
$$\left(Q_1\square^+_{n_1,n_2}Q_3\right)\nabla^+(Q_2\square^+_{n_1,n_2}Q_3). \blacksquare$$

Таким образом, в рамках модифицированной модели за счет расширения множества возвращаемых фрагментов обеспечивается ассоциативность операции нахождения последовательности фрагментов, а также дистрибутивность операции $\nabla^+$ относительно операций $\Delta^+_{n_1}$ и $\square^+_{n_1,n_2}$.

### 2.6 Комбинированная модель правил

Следующее утверждение показывает, что модифицированные правила можно использовать совместно с правилами для редуцированных множеств.

**Утверждение 8.** Пусть имеются произвольные правила $Q_1$ и $Q_2$, $\odot$ обозначает одну из операций $\nabla$, $\Delta$, $\square$, $\Delta_{n_1}$ и $\square_{n_1,n_2}$, а $\oplus$ обозначает одну из операций $\nabla^+$, $\Delta^+$, $\square^+$, $\Delta^+_{n_1}$ и $\square^+_{n_1,n_2}$, тогда справедливо следующее соотношение

$$F_{Q_1\odot Q_2} = R\left(F_{Q_1\oplus Q_2}\right). \blacksquare$$

Данное свойство позволяет построить комбинированную модель, в которой объединяются правила для двух рассмотренных моделей следующим образом:

1. Элементарным правилом будем называть правило $Q = t$, $t \in T$, результат которого $F_Q$ – редуцированное множество фрагментов, полученное в результате выполнения одной операции.

2. Модифицированным сложным правилом будем называть такое правило $Q$, которое получено путем выполнения одного из правил $\nabla^+$, $\Delta^+$, $\square^+$, $\Delta^+_{n_1}$, $\square^+_{n_1,n_2}$, $\wr$, $\lhd$, $\rhd$ над другими правилами $Q_1, \dots, Q_k$, которые являются либо элементарными правилами, либо модифицированными сложными правилами.

2. Сложным правилом будем называть такое правило $Q$, которое получено путем выполнения операций $\nabla$, $\Delta$, $\square$, $\bowtie$, $\wr$, $\lhd$, $\rhd$, $\Delta_{n_1}$, $\square_{n_1,n_2}$ над другими правилами $Q_1, \dots, Q_k$, которые являются либо элементарными, либо сложными, либо модифицированными сложными правилами с выполненной редукцией.

Таким образом, в данной модели модифицированные правила используются только в случае, когда задаются ограничения на расстояния между фрагментами. При этом при необходимости выполняется редукция результатов модифицированных правил.

## 3. Реализация правил классификации

### 3.1 Реализация правил для редуцированных множеств фрагменов

В алгоритмах, реализующих операции над редуцированными множествами фрагментов на входе и выходе используются редуцированные множества, которые дополнительно упорядочены с помощью отношения $\ll$. Для каждого такого алгоритма проведено формальное доказательство корректности, а также оценка вычислительной сложности. Из-за ограниченного объема работы приведем только полученные оценки вычислительной сложности.

**Утверждение 9.** Пусть имеется правила $Q_1$ и $Q_2$, результаты которых упорядоченные не пустые редуцированные множества фрагментов $F_{Q_1} = A = \left(a_1, \dots, a_{n_A}\right)$ и $F_{Q_2} = B = \left(b_1, \dots, b_{n_B}\right)$, где $a_1 \ll a_2 \ll \dots \ll a_{n_A}$, $b_1 \ll b_2 \ll \dots \ll b_{n_B}$, $\odot$ одно из правил $\nabla$, $\Delta$, $\square$, $\bowtie$, $\wr$, $\lhd$, $\rhd$, $\Delta_{n_1}$, $\square_{n_1,n_2}$. Тогда существует алгоритм, который строит редуцированное упорядоченное множество фрагментов $C = F_{Q_1\odot Q_2}$, где $C = (c_1, \dots, c_{n_C})$, $c_1 \ll \dots \ll c_{n_C}$, $n_C \le n_A + n_B$ и имеет вычислительную сложность порядка $O(n_A + n_B)$. ■

**Замечание.** В приведенном утверждении требуется упорядоченность исходных множеств фрагментов. Для элементарных правил данное требование обеспечивается за счет создания на этапе предварительной обработки документов специального индекса, в котором для каждого термина сохраняется упорядоченный набор его координат в тексте, который является упорядоченным редуцированным множеством. Для сложных правил данное требование выполняется по построению.

В условиях приведенного замечания и хранении элементов документа в виде упорядоченного списка, в котором выполняется бинарный поиск, справедливо следующее утверждение.

**Утверждение 10.** Пусть имеется сложное правило $Q$, которое включает $k$ элементарных

правил ($n_i$ - число фрагментов, возвращаемое правилом $i = 1, \dots, k$), и документ $D$ длины $n$, который содержит $m$ различных элементов (терминов), тогда вычислительная сложность нахождения $F_Q$ не более $O\left(k \log m + k\sum_{i=1,..k} n_i\right)$ операций ■.

Таким образом, вычислительная сложность не зависит от длины документа, а зависит только от количества элементарных запросов и числа соответствующих им элементов в словаре, что делает нахождение фрагментов более эффективным по сравнению с методами, основанными на использовании регулярных выражений.

### 3.2 Реализация правил для модифицированной модели

Для модифицированной модели множества фрагментов на входе и на выходе алгоритмов реализации правил являются частично редуцированными. В худшем случае данные множества могут содержать до $n(n+1)/2$ элементов, где $n$ число элементов в документе, что делает построение данных множеств в общем случае более затратным.

Свойства алгоритмов для вычисления данных правил описываются следующим утверждением.

**Утверждение 11.** Пусть имеется правила $Q_1$ и $Q_2$, результаты которых не пустые множества фрагментов $F_{Q_1} = A = \left(a_1, \dots, a_{n_A}\right)$ и $F_{Q_2} = B = \left(b_1, \dots, b_{n_B}\right)$, $\oplus \in \left\{\nabla^+, \Delta^+, \square^+, \Delta^+_{n_1}, \square^+_{n_1,n_2}\right\}$. Тогда существует алгоритм, который строит частично редуцированное множество фрагментов $C = F_{Q_1 \oplus Q_2}$, где $C = (c_1, \dots, c_{n_C})$, $n_C \le n_A n_B$ и имеет вычислительную сложность порядка $O(n_A n_B)$. ■

**Замечание.** Можно показать, что для операции $\nabla^+$ вычислительная сложность составляет порядка $O(n_A + n_B)$ операций.

Таким образом, вычислительная сложность правил для данной модели значительно выше сложности правил для редуцированной модели, что ограничивает возможности их прямого практического использования.

### 3.3 Релизация правил для комбинированной модели

Для реализации комбинированной модели необходимо выполнение редукции множеств, получаемых в результате правил для модифицированной модели. Следующее утверждение дает оценку вычислительной сложности для данной операции при использовании алгоритма «карманной» сортировки.

**Утверждение 12.** Пусть имеется множество фрагментов $A = \left(a_1, \dots, a_{n_A}\right)$, тогда для его редукции достаточно $O(n_A)$ операций и $O(n)$ памяти. ■

Выполнение редукции результатов выполнения модифицированных сложных правил позволяет организовать более эффективное их вычисление с использованием следующего подхода.

**Определение.** Будем говорить, что множество фрагментов $A \subset \mathbb{F}$ редуцировано справа, если для любого $a = (a_l, a_r) \in A$ не существует $a' = (a'_l, a'_r) \in A$ такого, что $(a'_l = a_l) \wedge (a'_r < a_r)$ .

**Определение.** Будем говорить, что множество фрагментов $A \subset \mathbb{F}$ редуцировано слева, если для любого $a = (a_l, a_r) \in A$ не существует $a' = (a'_l, a'_r) \in A$ такого, что $(a'_r = a_r) \wedge (a'_l > a_l)$ .

Операции левой и правой редукции произвольного множества $A$ будем обозначать $R_l(A)$ и $R_r(A)$. Свойства данных множеств описываются следующим утверждением.

**Утверждение 13.**

1. Пусть множество фрагментов $A \subset \mathbb{F}$ редуцировано слева или справа, тогда оно содержит не более $n$ элементов.

2. Пусть имеется произвольное множество фрагментов $A = (a_1, \dots, a_{n_A})$, тогда для его левой (правой) редукции требуется не более $O(n_A)$ операций.

3. Пусть $Q_1$ и $Q_2$ произвольные правила, результат выполнения которых множества $F_{Q_1}$ и $F_{Q_2}$, соответственно. Тогда справедливы следующие соотношения:

$R\left(F_{Q_1 \oplus Q_2}\right) = F_{Q'_1 \odot Q'_2}$, где $F_{Q'_1} = R(F_{Q_1})$, $F_{Q'_2} = R(F_{Q_2})$, $\oplus \in \{\nabla^+, \Delta^+, \square^+\}$ и $\odot \in \{\nabla, \Delta, \square\}$;

$R\left(F_{Q_1 \square^+_{n_1,n_2} Q_2}\right) = R\left(F_{Q'_1 \square^+_{n_1,n_2} Q'_2}\right)$, где $F_{Q'_1} = R_l(F_{Q_1})$, $F_{Q'_2} = R_r(F_{Q_2})$;

$R_x\left(F_{Q_1 \oplus Q_2}\right) = R_x(F_{Q'_1 \oplus Q'_2})$, где $F_{Q'_1} = R_x(F_{Q_1})$, $F_{Q'_2} = R_x(F_{Q_2})$, $x \in \{l, r\}$, $\oplus \in \{\nabla^+, \Delta^+, \square^+\}$ и $\odot \in \{\nabla, \Delta, \square\}$;

$R_l\left(F_{Q_1 \square^+_{n_1,n_2} Q_2}\right) = R_l\left(F_{Q'_1 \square^+_{n_1,n_2} Q_2}\right)$, где $F_{Q'_1} = R_l(F_{Q_1})$ и $R_r\left(F_{Q_1 \square^+_{n_1,n_2} Q_2}\right) = R_r\left(F_{Q_1 \square^+_{n_1,n_2} Q'_2}\right)$, где $F_{Q'_2} = R_r(F_{Q_2})$;

4. Если одно из входных множеств фрагментов $A = \left\{a_1, \dots, a_{n_A}\right\}$ или $B = \left\{b_1, \dots, b_{n_B}\right\}$ у операции $\square^+_{n_1,n_2}$ является редуцированным справа или слева, то вычислительная сложность данного правила не более $O(n_A + n_B)$. ■

Выполнение левой и правой редукции результатов промежуточных операций позволяет существенно снизить вычислительную сложность модифицированных операций как за счет уменьшения размера входных множеств, так и результирующего множества фрагментов. При этом максимальная сложность правила в комбинированной модели определяется количеством вложенных операций с ограничениями на расстояния.

## 4. Заключение

В работе рассмотрены модели задания правил классификации на основе операции с множествами

фрагментов. Данный подход обладает следующими преимуществами:

- позволяет по результатам классификации текста выделять в нем не отдельные слова, а значимые фрагменты, которые полностью соответствуют сложным условиям правила;
- позволяет задавать ограничения на расстояния не только между отдельными словами, а между сложными выражениями;
- позволяет реализовать эффективные алгоритмы выделения фрагментов при использовании множеств фрагментов специального вида (сложность линейно зависит от размера правила и не зависит от длины текста).

Проведен анализ алгебраических и вычислительных свойств различных операций по результатам которого предложена комбинированная модель представления правил, которая, с одной стороны, обеспечивает ассоциативность и дистрибутивность основных операций, а с другой стороны, является достаточно эффективной.

На основе приведенных в настоящей работе подходов и алгоритмов в рамках системы СКАТ разработан специальный язык для задания правил классификации текстов, который также имеет следующие основные возможности:

- задание ограничений на разделы текста и размер области, в которой должны находиться термины или сложные выражения;
- задание ограничений на частоту встречаемости и веса терминов в различных разделах текста;
- задание ограничений на морфологические и графематические характеристики терминов;
- задание ограничений на встречаемость терминов с учетом и без учета морфологии, наличия ошибок или похожести по звучанию;
- определение сложных понятий и шаблонных выражений, которые можно одновременно использовать в различных правилах классификации;
- задание специальных правил, учитывающих иерархическую структуру дерева рубрик и результаты выполнения правил в других рубриках;
- задание комбинированных правил, в которых фрагменты могут выделяться и с использованием статистических методов, основанных на обучении на примерах.

С использованием данного языка успешно решались задачи по географической классификации текстов по субъектам РФ, по построению различных тематических классификаторов и выделению описаний различных типов ситуаций и объектов в текстах.

К перспективным направлениям дальнейших исследований можно отнести следующие: разработка методов автоматической коррекции правил классификации за счет оценки документов пользователями; разработка подходов к объединению результатов выделения фрагментов в различных рубриках; разработка эффективных подходов к интеграции подходов к классификации на основе правил и обучения на примерах.

## Литература


[1] Clarke C.L.A, Cormack G.V. Shortest-Substring Retrieval and Ranking // ACM Transactions on Information Systems, Vol. 18, No. 1, January 2000, pp. 44-78.

[2] Dumais S.T., Lewis D.D. & Sebastiani F., Report on the Workshop on Operational Text Classification Systems // SIGIR-02, Tampere, Finland. – 4 p. (http://www.sigir.org/forum/F2002/sebastiani.pdf).

[3] Dominich S. The Modern Algebra of Information Retrieval. The Information Retrieval Series. – Berlin, Springer-Verlag, 2008. – 327 p. (ISBN 978-3-540-77658-1)

[4] [HW90] Hayes P. J., Weinstein S. P. Construe-TIS: A System for Content-Based Indexing of a Database of News Stories // IAAI-90 Proceedings, 1990. – 16 p.

[5] Manning C.D., Raghavan P., Schutze H. An Introduction to Information Retrieval. – Cambridge University Press, 2008. – 577

[6] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.

[7] Агеев М.С., Добров Б.В., Лукашевич Н.В. Поддержка системы автоматического рубрицирования для сложных задач классификации текстов // Труды 6-ой Всероссийской научной конференции – RCDL2004, Пущино, Россия, 2004. – 10 с.

[8] Васильев В.Г. Комплексная технология автоматической классификации текстов // Компьютерная лингвистика и интеллектуальные технологии: по материалам конференции «Диалог». Вып. 7(14). – М. РГГУ, 2008. - с. 83-90.

[9] Васильев В.Г. Выделение фрагментов в текстах при классификации // Компьютерная лингвистика и интеллектуальные технологии: по материалам конференции «Диалог». Вып. 8(15). – М. РГГУ, 2009. - с. 57-63.


## Fragment Extraction and Text Classification by Logical Rules

© V.G. Vasilyev

In this paper a method of text classification based on using formal mathematical model of operations on fragment sets is described. Algebraic properties of different definitions of rules are analyzed. Particular attention is paid to rules with restrictions on distance and relative positions of fragments. Efficient algorithms for rules are developed and justified.