УДК 681.3

А.М. Наместников

КОНЦЕПТУАЛЬНОЕ ИНДЕКСИРОВАНИЕ ПРОЕКТНЫХ ДОКУМЕНТОВ НА ОСНОВЕ ГЕНЕТИЧЕСКОЙ ОПТИМИЗАЦИИ[1]

Наместников Алексей Михайлович, *кандидат технических наук, доцент, окончил радиотехнический факультет Ульяновского государственного технического университета. Доцент кафедры «Информационные системы» факультета информационных систем и технологий. Имеет статьи и монографию в области интеллектуальных систем хранения и обработки информации. [e-mail: nam@ulstu.ru].*

Аннотация

В статье содержится формальное описание процесса нахождения доминирующего понятия в текстовом фрагменте проектного документа (ПД). Представлена адаптация стандартного генетического алгоритма для решения задачи оптимального разбиения проектного документа на текстовые фрагменты с доминированием понятий онтологии. Рассмотрены реализации операторов кроссинговера и мутации.

Ключевые слова: интеллектуальная система, индексирование, генетический алгоритм, онтология.

Alexey Mikhailovich Namestnikov, *Candidate of Engineering, Associate Professor; graduated from the Faculty of Radioengineering of Ulyanovsk State Technical University; Associate Professor at the Chair 'Information Systems' of the Faculty of Information Systems and Technology; author of articles and a monograph in the field of intelligent systems for storage and processing of information. e-mail: nam@ulstu.ru.*

Abstract

The article contains a formal description for a process of finding of a dominant concept in a text fragment of a design document. It also presents an adaptation for standard genetic algorithm in order to solve the task of optimum segmentation of design document into text fragments, ontology concepts dominated, and considers implementations of crossingover and mutation statements.

Key words: intellectual system, indexing, genetic algorithm, ontology.

Введение

Производя анализ содержимого любого ПД, можно сделать вывод о том, что, несмотря на единое общее стилевое решение, в различных его частях (фрагментах) составитель такого документа делает акцент на разных понятиях предметной области. В этой связи возникает задача учитывать данный факт в процессе концептуального индексирования ПД, опираясь при этом на описание предметной области в виде онтологии. В данной работе будем исходить из предположения о том, что текст анализируемого документа уже прошел предобработку и представляет собой последовательность термов, являющихся элементами предложений документа. Кроме того, будем считать минимальным фрагментом анализируемого текста отдельное предложение, а максимальным – текстовый документ в целом.

1 Определение доминирующего понятия в текстовом фрагменте проектного документа

Пусть имеется предобработанный текстовый документ d, состоящий из последовательности термов:

$$S^d = w^d_{11}, w^d_{21}, \ldots, w^d_{i_1 1}, \ldots, w^d_{n_1 1}, w^d_{12}, \ldots, w^d_{i_2 2}, \ldots, w^d_{n_2 2}, \ldots, w^d_{i_j j}, \ldots, w^d_{n_m m}, \quad (1)$$

где i_j – номер терма в k-м предложении, $j = \overline{1, m}$;

$i_j = \overline{1, n_j}$, где n_j – количество термов в j-м предложении.

Онтология предметной области включает в себя два уровня: концептуальный и терминологический [1, 2]. Концептуальный уровень представляется в виде дерева

$$(C, E),$$

где $C = \{c\}$ – множество концептов (понятий) предметной области, зафиксированных в онтологии;

$E = \{<c_i, c_k>\}, <c_i, c_k> \in C^2$ – множество дуг, соединяющих понятия.

Терминологический уровень для k-го понятия записывается в виде множества:

$$\{(w^k_1, f^k_1), (w^k_2, f^k_2), \ldots, (w^k_i, f^k_i), \ldots, (w^k_{l_k}, f^k_{l_k})\},$$

где w^k_i – i-й терм k-го понятия онтологии;

l_k – общее количество термов, ассоциированных с k-м понятием;

1 Работа выполнена при финансовой поддержке РФФИ, проект №10-07-00064-а.

f_i^k – частота встречаемости i-го терма в описании k-го понятия.

Обозначим через S_p^d часть последовательности S^d, которая определяется выражением (1), и запишем ее следующим образом:

$$S_p^d = w_{1p}^d, w_{2p}^d, \dots, w_{j_p p}^d, \dots, w_{n_p p}^d,\ p = \overline{1,s},$$

при этом выполняется равенство:

$$S_1^d, S_2^d, \dots, S_s^d = S^d. \quad (2)$$

В процессе концептуального индексирования ПД необходимо определить набор понятий предметной области, который содержится в тексте анализируемого документа. Будем принимать во внимание следующую гипотезу.

Гипотеза: *Любой текстовый документ можно разделить на множество непересекающихся фрагментов, в каждом из которых будет доминировать тот или иной концепт предметной области.*

Для нахождения значения доминирования концептов будем применять предложенный в работе [3] метод сравнения текстового входа каждого понятия в онтологии предметной области с анализируемым текстом.

Степень выраженности понятия c_k в p-м фрагменте ПД d будем вычислять по следующей формуле:

$$\mu_{S_p^d}(c_k) = 1 - \frac{1}{l_k}\sum_{s=1}^{l_k}\left|f_s^k - f_s^p\right|,$$

где S_p^d – p-й фрагмент ПД d;

f_s^p, f_s^k – частоты встречаемости термина s в p-м фрагменте документа и в описании k-го понятия онтологии соответственно;

l_k – мощность текстового входа понятия c_k.

В том случае, если термин s отсутствует в p-м фрагменте документа d, f_s^p принимается равным нулю.

На рисунке 1 представлены две диаграммы степеней выраженности понятий предметно-ориентированной онтологии в текстовых фрагментах S_1 и S_2. Предполагая, что выделение из документа текстового фрагмента преследовало цель ограничения некоторого концепта, можно утверждать, что в рамках фрагмента S_1 (рис. 1а) данная операция была выполнена более успешно, чем во фрагменте S_2 (рис. 1б). На рисунке 1а явно наблюдается доминирование концепта c_1 относительно других концептов по степени его выраженности в текстовом фрагменте S_1, что отсутствует во фрагменте S_2.

Алгоритм вычисления степени доминирования понятия в текстовом фрагменте состоит из следующих шагов:

Шаг 1. Определение максимальной степени выраженности концептов в текстовом фрагменте (рис. 2):

$$\hat{\mu}_{S_p^d}(c) = \max_c\left(\mu_{S_p^d}(c)\right).$$

Шаг 2. Определение среднего значения степени выраженности концептов онтологии, исключая концепт с максимальной степенью выраженности (определенный на предыдущем шаге):

$$\tilde{\mu}_{S_p^d}(c) = \frac{1}{n-1}\sum_{i=1}^{n-1}\mu_{S_p^d}(c_i),$$

где $c_i \in c - c_k$; $c_k = \arg\max_c\left(\mu_{S_p^d}(c)\right)$;

n – количество концептов с ненулевой степенью выраженности для текстового фрагмента S_p^d.

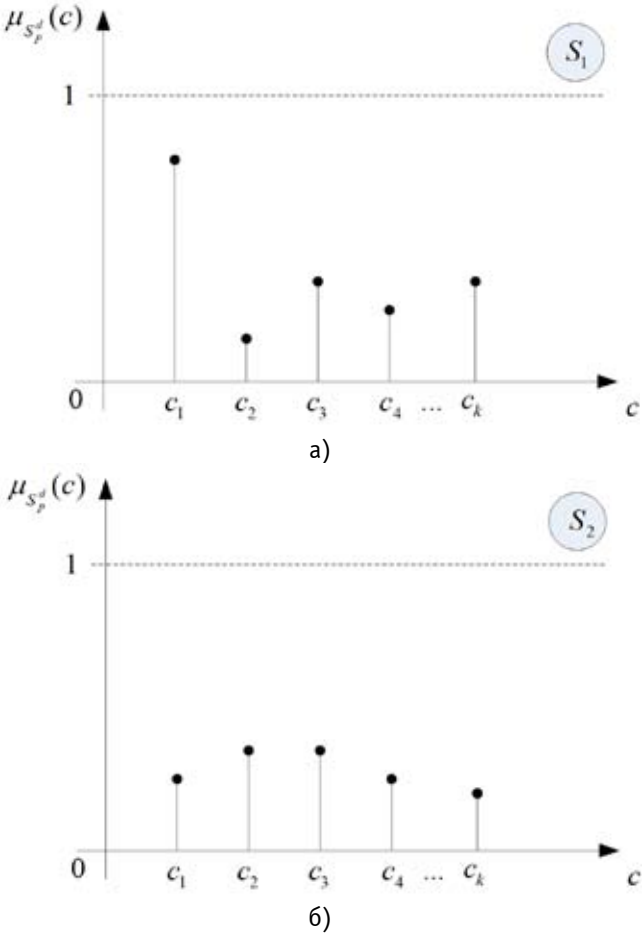


Рис. 1. Диаграммы степеней выраженности понятий онтологии в текстовом фрагменте

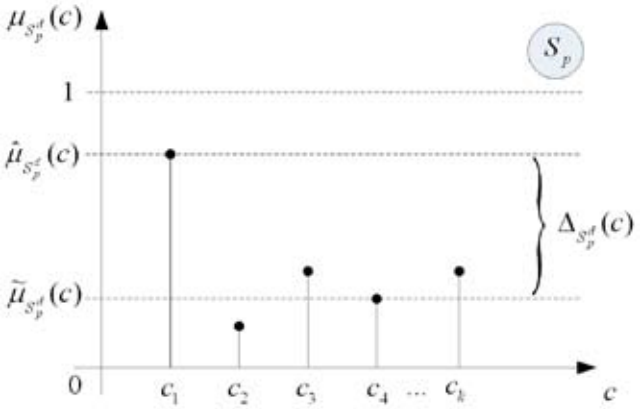


Рис. 2. Определение степени детерминирования понятия в текстовом фрагменте

Шаг 3. Определение степени детерминированности понятия в текстовом фрагменте S_p^d:

$$\Delta_{S_p^d}(c) = \hat{\mu}_{S_p^d}(c) - \tilde{\mu}_{S_p^d}(c). \quad (3)$$

Выражение (3) фактически определяет качество выделения текстового фрагмента в ПД с целью ограничения в тексте определенного понятия предметной области, которое зафиксировано в онтологии интеллектуального проектного репозитория.

2 Адаптация генетического алгоритма для задачи концептуального индексирования

Целью генетической оптимизации в процессе концептуального индексирования ПД является нахождение такой последовательности (2), которая соответствует минимальному значению целевой функции:

$$F(S^d) = \sum_p (1 - \Delta_{S_p^d}(c)) \to \min, \quad (4)$$

$p = \overline{1, s}$, где s – количество текстовых фрагментов, $s = \overline{1, m}$,

где m – количество предложений в индексируемом документе.

Таким образом, минимальный текстовый фрагмент соответствует одному отдельно взятому предложению ПД, а максимальный – целому ПД.

Канонический генетический алгоритм (рис. 3) характеризуется следующими особенностями [4]:

1) Задается функция оптимальности (она же – целевая функция), определяющая эффективность найденного решения.

2) В соответствии с определенными ограничениями инициализируется исходная популяция потенциальных решений. Каждое решение кодируется как вектор, который называется хромосомой. Его элементами являются части вектора – гены, у которых изменяющиеся значения в определенных позициях называются аллелями.

3) Каждая хромосома в популяции декодируется в необходимую форму для последующей оценки, затем ей присваивается значение эффективности в соответствии с целевой функцией.

4) Каждой хромосоме присваивается вероятность воспроизведения, которая зависит от эффективности данной хромосомы.

Рис. 3. Блок-схема генетического алгоритма

5) В соответствии с вероятностями воспроизведения создается новая популяция хромосом, причем с большей вероятностью воспроизводятся наиболее эффективные элементы. Хромосомы производят потомков, используя операции рекомбинации: кроссинговер (хромосомы скрещиваются, обмениваясь частями строк) и мутация (вероятностное изменение аллелей).

Формально генетический алгоритм можно описать следующим образом [4]:

$$ГА = (P^0, \lambda, l, \nu, \rho, F, \tau),$$

где $P^0 = (a_1^0, ..., a_\lambda^0)$ – исходная популяция, где a_i^0 – решение задачи, представленное в виде хромосомы;

λ – целое число (размер популяции);

l – целое число (длина каждой хромосомы популяции);

ν – оператор отбора;

ρ – отображение, определяющее рекомбинацию (кроссинговер, мутация);

F – целевая функция;

τ – критерий остановки.

Для решения конкретной задачи оптимизации текстовых фрагментов ПД генетический алгоритм требует следующих уточнений:

- способа кодирования хромосом (потенциальных решений);
- вида целевой функции;
- реализации операций кроссинговера и мутации.

Потенциальное решение (хромосома) генетического алгоритма концептуального индексатора имеет следующий вид:

$$a_i^t = (< p, j >), \qquad p = \overline{1, s},$$

$$j = \overline{1, m}, \; 1 \le s \le m, \quad (5)$$

где p – номер текстового фрагмента;

j – номер предложения;

s – количество текстовых фрагментов;

m – количество предложений;

i – номер хромосомы;

t – номер поколения.

Таким образом, хромосома, определяемая выражением (5), представляет из себя, в действительности, последовательность текстовых фрагментов (2).

Целевая функция определяет способ отображения хромосомы на единичный отрезок:

$$F : a_i^t \to [0,1].$$

В качестве целевой функции F будем использовать выражение (4).

На первом шаге работы генетического алгоритма формируется начальная популяция хромосом $P^0 = (a_1^0,...,a_\lambda^0)$. Для каждой хромосомы a_i^0 определяется значение целевой функции $F(a_i^0)$. Затем производится ранжирование хромосом. Ранг элементов популяции $rank$ задается следующим образом:

$$\forall i \in \{1,...,\lambda\} : rank(a_i^t) = i,$$

если для $\forall j \in \{1,\dots,\lambda-1\} : F(a_j^t) < F(a_{j+1}^t)$.

Первые g хромосом без изменения переходят в следующий пул (поколение), а остальное количество $(\lambda - g)$ формируется посредством операции кроссинговера. При определении оператора кроссинговера будем учитывать то, что последовательность предложений в тексте и их количество должны оставаться неизменными в процессе трансформации хромосом. Точка кроссинговера определяется случайным образом на границе двух текстовых фрагментов:

$$a_i^0 = (...,< p, j >,< p+1, j+1 >,...)$$

для первой из двух хромосом, участвующих в кроссинговере. Так как в процессе рассматриваемой операции происходит взаимообмен частями хромосом и принимая во внимание вышеприведенные ограничения, точку кроссинговера для второй хромосомы выбираем так, чтобы в левой части остались j первых предложений, как и у первой хромосомы. Поскольку принцип разбиения на текстовые фрагменты для указанных хромосом может быть различным (с точностью до предложения), то нельзя исключать ситуацию, когда в процессе кроссинговера во второй хромосоме будет добавлен еще один текстовый фрагмент. Это возможно по причине сохранения равного количества предложений в хромосомах, и, как следствие, всегда имеется ненулевая вероятность разбиения одного текстового фрагмента второй хромосомы на два фрагмента. Иллюстративно данная ситуация показана на рисунке 4.

Отбор хромосом для кроссинговера производится на основании вероятностей $p_v(a_i^t)$, вычисленных для каждого индивидуума популяции с использованием метода пропорционального отбора:

$$p_v(a_i^t) = F(a_i^t) / \sum_{j=1}^{\lambda} F(a_j^t).$$

Заключительным этапом формирования новой популяции является применение оператора мутации. В задаче концептуального индексирования предлагается применять два варианта мутации хромосом: 1) сдвиг границы текстового фрагмента; 2) объединение текстовых фрагментов (рис. 5).

Первый вариант мутации со сдвигом границы текстового фрагмента предполагает вероятностный выбор границы между двумя текстовыми фрагментами ПД. Далее принимается решение о направлении сдвига границы в правую или в левую сторону, учитывая количество предложений в соседних текстовых фрагментах. Граница перемещается на одно предложение в сторону с большим

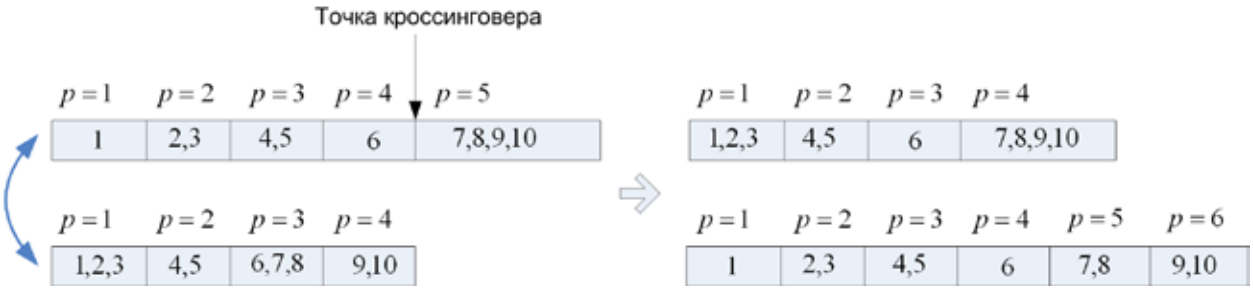
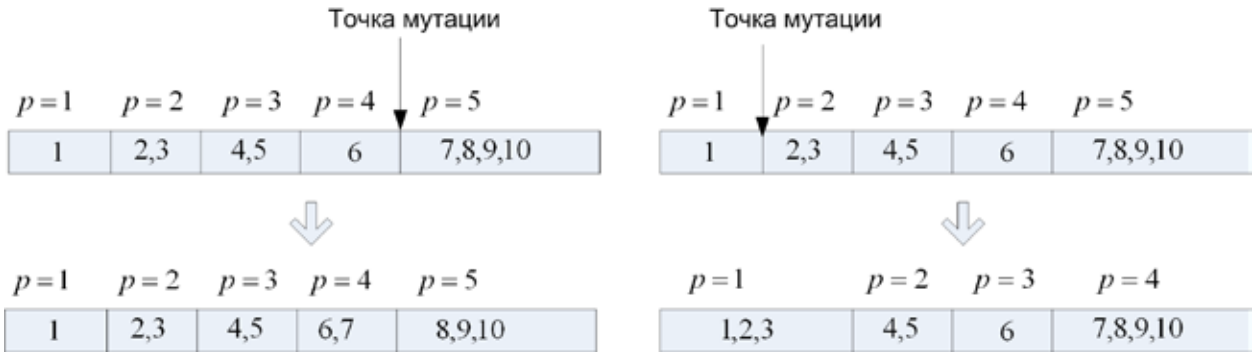


Рис. 4. Иллюстративный пример операции кроссинговера



а) мутация со сдвигом границы

б) мутация объединения

Рис. 5. Иллюстративный пример оператора мутации

количеством предложений. При равенстве предложений направление выбирается случайным образом. Сдвиг границы не происходит в случае, если текстовый фрагмент содержит одно предложение. На рисунке 5а показан пример оператора мутации, который осуществляет сдвиг границы между четвертым и пятым текстовыми фрагментами вправо.

Вариант мутации, объединяющий два соседних фрагмента, фактически уменьшает количество текстовых фрагментов в ПД за счет их укрупнения. На рисунке 5б представлен пример оператора мутации, выполняющий объединение первого и второго текстовых фрагментов.

В процессе настройки параметров генетического алгоритма необходимо правильным образом подобрать значения вероятностей мутации первого и второго вариантов для того, чтобы:

- обеспечивалась удовлетворительная сходимость генетического алгоритма;
- происходило формирование достаточного разнообразия вариантов хромосом (с целью выхода или непопадания в локальные экстремумы целевой функции);
- обеспечивался баланс между ростом количества текстовых фрагментов в процессе кроссинговера и их уменьшением благодаря оператору мутации с объединением текстовых фрагментов.

Заключение

В данной работе представлено формальное описание доминирования концепта в произвольном текстовом фрагменте ПД. Для нахождения оптимальных текстовых фрагментов с точки зрения описания предметной области в виде онтологии предлагается использовать генетический алгоритм. Произведена адаптация стандартного генетического алгоритма к задаче оптимального разбиения документа на текстовые фрагменты, заключающаяся в формировании целевой функции, реализации операторов кроссинговера и мутации и кодировании хромосом (потенциальных решений).

СПИСОК ЛИТЕРАТУРЫ

1. Наместников А.М., Чекина А.В., Корунова Н.В. Интеллектуальный сетевой архив электронных информационных ресурсов // Программные продукты и системы. – 2007. – № 4. – С. 10–13.

2. Наместников А.М., Филиппов А.А. Концептуальная индексация проектных документов //Автоматизация процессов управления. – 2010. – № 2 (20). – С. 34–39.

3. Наместников А.М., Филиппов А.А. Реализация системы кластеризации концептуальных индексов проектных документов // Автоматизация процессов управления. – 2011. – № 3 (25). – С. 46–50.

4. Скурихин А.Н. Генетические алгоритмы // Новости искусственного интеллекта. – 1995. – № 4. – С. 6–17.