



## КОМПЛЕКСНАЯ МЕТОДИКА ВЫЯВЛЕНИЯ ЗАИМСТВОВАНИЙ В ПИСЬМЕННЫХ УЧЕБНЫХ РАБОТАХ

**Шостак И. В., Груздо И. В.**

*Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ»,  
г. Харьков, ул Чкалова, тел. 788-47-35,  
e-mail: iv\_shostak@rambler.ru., tigralwovna@rambler.ru.*

Современный этап развития сферы образования в Украине характеризуется, с одной стороны, стремительным ростом объемов письменных работ, а с другой - расширением их номенклатуры. Проверка этих работ нуждается в тщательном анализе со стороны преподавателей с целью установления оригинальности контента, а также в одновременном использовании при оценивании работ четырех разнотипных систем. Указанные обстоятельства порождают проблему, суть которой состоит в недостаточной эффективности существующих, «ручных» способов обработки преподавателями текстов письменных учебных работ (ПУР), установлении фактов наличия в них текстовых заимствований, и формировании на основании этих фактов объективных оценок работ в четырех разнотипных системах оценивания.

Для создания методики поиска заимствований в ПУР важным вопросом является выбор методов и языков формального описания. При этом необходимо обеспечить выполнение следующих требований: математической строгости; наглядности представления; отсутствия избыточности; легкости понимания; возможности описания, как алгоритмов функционирования, так и управления процессами анализа текста; наличие специального инструментария для реализации соответствующей информационной технологии.

Анализ проведенных в [1, 2] методов поиска текстовых заимствований показал, что выбор определенного из них, зависит от значимости рубрик ПУР. Поэтому в начале необходимо разработать процессную модель выбора метода поиска заимствований. На следующем этапе необходимо определить метод поиска заимствований в тексте, наиболее подходящий для конкретного типа ПУР с учетом введенных ограничений. Входными данными в этом случае будут целевые функции методов выбора, критерием выбора – тип и объем ПУР, при этом учитывается номенклатура рубрик присущих данному виду ПУР. Ограничениями являются вид поиска (точный, при котором находятся все посимвольные совпадения строки в тексте с заданным образцом; или нечеткий, при котором в искомой строке могут встречаться несоответствия с заданным образцом). Результатом работы модели является применение в ходе поиска последовательности методов, оптимальной для поиска заимствований в конкретном типе ПУР.

Обозначим  $X$  множество альтернатив, т.е. возможных вариантов;  $K$  – тип ПУР;  $S$  – виды поиска (0 - нечеткий поиск и 1 - точный поиск);  $V_s$  – объем ПУР;  $R$  – наличие всех (1) или обязательных рубрик (0) для конкретного типа ПУР.



В зависимости от требований и условий, выбираем метод для выполнения поиска заимствований. Для наглядности представим формально критерии выбора в таблице 1.

Таблица 1 – Формальное представление критериев выбора методов поиска для различных типов ПУР

Метод	Тип ПУР(К)	Рубрики (R)	Объем ПУР(V)	Тип поиска(S)
1 Рабина	$ПУР_{лаб}$	R=1	$10 < V < 16$	S=1
	$ПУР_{реф}$	R=0	$10 < V < 16$	S=1
	$ПУР_{дз}$	R=0	$10 < V < 16$	S=1
2 Бойера-Мура-Хорспула	$ПУР_{дз}$	R=1	$15 < V < 60$	S=1
	$ПУР_{реф}$	R=1	$10 < V < 60$	S=1
	$ПУР_{ер}$	R=0	$50 < V < 60$	S=1
3 Маасса-Новака	$ПУР_{ер}$	R=1	$60 < V < 120$	S=1
	$ПУР_{дн}$	R=0	$60 < V < 120$	S=1
	$ПУР_{ор}$	R=0	$60 < V < 120$	S=1
4 Левенштейна	$ПУР_{лаб}$	R=1	$10 < V < 16$	S=0
	$ПУР_{кп}$	R=0	$10 < V < 16$	S=0
	$ПУР_{дз}$	R=0	$10 < V < 16$	S=0
	$ПУР_{кр}$	R=0	$10 < V < 16$	S=0
5 Ханта-Шиманского	$ПУР_{дз}$	R=1	$15 < V < 50$	S=0
	$ПУР_{кп}$	R=1	$10 < V < 40$	S=0
	$ПУР_{кр}$	R=1	$10 < V < 40$	S=0
	$ПУР_{ер}$	R=0	$50 < V < 60$	S=0
6 Машека-Патерсона	$ПУР_{ер}$	R=1	$60 < V < 120$	S=0
	$ПУР_{дн}$	R=1	$60 < V < 120$	S=0
	$ПУР_{ор}$	R=1	$60 < V < 120$	S=0

Методика поиска заимствований в ПУР с учетом типа и значимости структурных частей ПУР предусматривает реализацию следующих этапов:

1) Преобразование анализируемой ПУР – в вид, удобный для построения полнотекстового индекса, вычисляемого автоматически для каждой рубрики (выделение из документов содержательной информационной основы).

2) Выполнение автоматической классификации анализируемой ПУР по матрице согласования приоритетов уровней, в которой поставлены в соответствие компоненты уровней иерархии ПУР и значения функции принадлежности, каждого уровня в зависимости от значимости структурных частей ПУР.



а) Вычисление признакового пространства  $P = P_1 \times P_2 \times \dots \times P_N$ ,  $i$ -го признака  $P_i$  рассматриваемой ПУР.

б) Нахождение признакового описания рассматриваемой ПУР  $d_i$  по заданному в матрице согласования приоритетов уровней.

в) Определение отношения  $\Phi : D \times C \rightarrow \{0,1\}$ , для каждой пары  $(d_i, c_j)$ , относится ли данная ПУР  $d_i$ , имеющая признаковое описание  $f(d_i)$ , к категории  $c_j$  для этого находим из матрицы максимальную близкую к функции  $\Phi$  функцию  $\Phi_{ПУР}$ . Каждое значение  $f(d_i)$ , нормируются от нуля до единицы. Поэтому и расстояние нормировано к единице.

3) Нахождение множества ПУР в банке данных работ, соответствующих типу анализируемой ПУР.

4) Выбор поискового инструментария из следующей номенклатуры методов поиска  $\alpha_1$  – Рабина;  $\alpha_2$  – Бойера-Мура-Хорспула;  $\alpha_3$  – Маасса-Новака;  $\alpha_4$  – Левенштейна;  $\alpha_5$  – Ханта-Шиманского;  $\alpha_6$  – Машека-Патерсона

5) Вычисление заимствования по выбранному методу с учетом значения условий перехода:

$$\alpha_1 = \begin{cases} 1 - \text{Если } X = (K = ПУР_{\text{лаб}}) \vee (R = 1) \vee (10 < V < 16) \vee (S = 1) \parallel \\ \quad (K = ПУР_{\text{реф}}) \vee (R = 0) \vee (10 < V < 16) \vee (S = 1) \parallel \\ \quad (K = ПУР_{\text{дз}}) \vee (R = 0) \vee (10 < V < 16) \vee (S = 1); \\ 0 - \text{если нет;} \end{cases}$$

$$\alpha_2 = \begin{cases} 1 - \text{Если } X = (K = ПУР_{\text{дз}}) \vee (R = 1) \vee (15 < V < 60) \vee (S = 1) \parallel \\ \quad (K = ПУР_{\text{реф}}) \vee (R = 1) \vee (10 < V < 60) \vee (S = 1) \parallel \\ \quad (K = ПУР_{\text{сп}}) \vee (R = 0) \vee (50 < V < 60) \vee (S = 1); \\ 0 - \text{если нет;} \end{cases}$$

$$\alpha_3 = \begin{cases} 1 - \text{Если } X = (K = ПУР_{\text{сп}}) \vee (R = 1) \vee (60 < V < 120) \vee (S = 1) \parallel \\ \quad (K = ПУР_{\text{дн}}) \vee (R = 0) \vee (60 < V < 120) \vee (S = 1) \parallel \\ \quad (K = ПУР_{\text{др}}) \vee (R = 0) \vee (60 < V < 120) \vee (S = 1); \\ 0 - \text{если нет;} \end{cases}$$

$$\alpha_4 = \begin{cases} 1 - \text{Если } X = (K = ПУР_{\text{лаб}}) \vee (R = 1) \vee (10 < V < 16) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{кн}}) \vee (R = 0) \vee (10 < V < 16) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{дз}}) \vee (R = 0) \vee (10 < V < 16) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{др}}) \vee (R = 0) \vee (60 < V < 120) \vee (S = 0); \\ 0 - \text{если нет;} \end{cases}$$

$$\alpha_5 = \begin{cases} 1 - \text{Если } X = (K = ПУР_{\text{дз}}) \vee (R = 1) \vee (15 < V < 50) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{кн}}) \vee (R = 1) \vee (10 < V < 40) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{кр}}) \vee (R = 1) \vee (10 < V < 40) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{сп}}) \vee (R = 0) \vee (50 < V < 60) \vee (S = 0); \\ 0 - \text{если нет;} \end{cases}$$



$$\alpha_5 = \begin{cases} 1 - \text{Если } X = (K = ПУР_{\text{сп}}) \vee (R = 1) \vee (60 < V < 120) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{оп}}) \vee (R = 1) \vee (60 < V < 120) \vee (S = 0) \parallel \\ \quad (K = ПУР_{\text{др}}) \vee (R = 1) \vee (60 < V < 120) \vee (S = 0); \\ 0 - \text{если нет.} \end{cases}$$

б) Формирование списка фрагментов заимствований из других текстов работ, при этом индекс каждого элемента списка будет указывать на структурную часть анализируемой ПУР.

Качественным отличием описанной методики анализа текстовых документов является учет особенностей, присущих письменным учебным работам, что, в конечном итоге, приводит к повышению эффективности организации учебного процесса в заведениях различного уровня аккредитации.

#### Список литературы

1. Груздо, I.B., Шостак, I.B. Анализ эффективности методов обнаружения плагиата в работах студентов технических специальностей [Текст] / I.B. Груздо, I.B. Шостак // Авіаційно-космічна техніка та технологія. Харків 2011. – 9(86). – С. 87.

2. Груздо, I.B., Шостак, I.B. Методи виявлення плагиату в роботах студентів технічного вузу [Текст] / I.B. Груздо, I.B. Шостак / XVI Міжнародний конгрес двигунобудівників 2011. - «ІКАР» п. Рибаче, АР Крим, 14-19 вересня 2011. - С. 88.