УДК 510.649:004.912

И.В. Груздо

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков

МОДЕЛЬ РУБРИЦИРОВАННОГО ОБЪЕКТА В ЗАДАЧАХ МАШИННОГО АНАЛИЗА ТЕКСТОВ, УЧИТЫВАЮЩАЯ ЗНАЧИМОСТЬ СТРУКТУРНЫХ ЧАСТЕЙ

Приведена формальная постановка задачи машинного анализа естественно языковых объектов с целью выявления заимствований на примере выявления плагиата в письменных работах обучающихся. Описана нечеткая структурная модель объекта текстологического анализа. Изложена методика определения расстояний между уровнями иерархий указанной модели.

Ключевые слова: *текстологический анализ, письменная работа, иерархическая нечеткая модель структуры текста, уровни иерархии*

Введение

На современном этапе развития информационных технологий как во всем мире, так и в Украине особую актуальность приобретают задачи, связанные с машинным поиском информации [1]. К этому классу задач относится и компьютеризация процессов анализа различного рода письменных документов с целью выявления в них заимствований. Одной из типичных задач указанного класса является автоматический поиск заимствований во всякого рода учебных работах.

Процесс выявления заимствований в учебных работах состоит из следующих этапов: определение специфики учебных работ, диагностика причин заимствования, классификация видов заимствований учебных работ, оценивание текстов учебных работ с позиций наличия в них плагиата, создание технологии выявления заимствований в учебных работах, организация машинного анализа учебных работ на предмет наличия в них плагиата с целью повышения качества учебного процесса в целом.

Реализация перечисленных выше этапов связана с решением еще одной важной прикладной задачи, а именно со структурированием анализируемой учебной работы, установлением семантических связей между отдельными структурными частями с учетом их значимости.

Целью статьи является описание специальной модели структуры текста в форме иерархической системы, которая дает возможность декомпозировать исходный текст на структурные части и поставить в соответствие каждой части уровень ее значимости

1. Постановка задачи исследования

Формально задача выявления текстологических заимствований, в различных типах учебных работ относится к классу задач принятия решений.

Пусть X-это конечное множество учебных работ, выполненных в данном ВУЗе, Q – конечное множество типов учебных работ, f - функция, отображающая $X \times Q$ в некоторое множество T, частично или полностью упорядоченное отношением $\le$. Пусть также заданна функция допустимости μ переводящая Q в T, т.е. $f: X \times Q \to T$ и $\mu: Q \to T$. Задача нахождения удовлетворяющих решений заключается в следующем: дано подмножество $X^m \subseteq X$, требуется найти такое μ из X^m, что для всех q из Q

$$f(x,q) \le \mu(q) \qquad (1)$$

$$\forall q_i \in Q \,|\, (q_i \in Q^{(Б)}) \vee (q_i \in Q^{(Ч)}), \quad i = \overline{1,N}.$$

Автоматическая классификация типов учебных работ является задачей распознавания образов с обучением. Сформулируем задачу. Пусть заданы: конечное множество категорий $C = \{c_1, c_2, \ldots, c_{|C|}\}$; конечное множество письменных учебных работ

$$D = \{d_1, d_2, \ldots, d_{|D|}\}; \qquad (2)$$

признаковое пространство $P = P_1 \times P_2 \times \ldots \times P_N$, где P_i – множество значений i-го признака; функция признака $f: D \to P$, $f(d_i) = (p_1, p_2, \ldots, p_N)$ – признаковое описание документа d_i. Имеется неизвестная функция $\Phi: D \times C \to \{0,1\}$, которая для каждой пары (d_i, c_j) определяет, относится ли данная учебная работа d_i, имеющая признаковое описание $f(d_i)$, к категории c_j; заданы: значения неизвестной функции Φ на некотором выбранном множестве учебных работ $D' \subset D$, D' – обучающее множество учебных работ; значения неизвестной функции Φ на некотором выбранном множестве учебных работ $D'' \subset D$ $(D' \cap D'' = \varnothing)$, D'' – тестовое множество работ.

Требуется найти максимально близкую к функции Φ функцию Φ', используя множество D', и оценить ее эффективность на множестве D''. Функцию Φ' называют классификатором.

Эта задача представляет собой задачу оптимизации с нечетко выраженными критериями и с большим числом возможных частных критериев, т.е. может быть отнесена к классу задач многокритериальной оптимизации, но при этом основная трудность, которая возникает при решении поставленной задачи, состоит в невозможности получения математического описания функции полезности U лица, принимающего решение.

Указанное выше обстоятельство определяет целесообразность использования теории нечетких множеств для математической формализации нечеткой информации. Иерархия декомпозиции рассматриваемой задачи представлена на рис/ 1.

2. Иерархическая модель учебной работы с учетом нечетких связей между структурными элементами

Обозначим множество видов учебных работ через G,а множество альтернатив через X. Если множества видов учебных работ отличны от рассматриваемых, их можно представить в виде иерархии, отражающей структурные части работы и связи между этими частями.

Элементы иерархи являются нечеткими множествами, обозначенными на рис. 2 как G_i, $i = 1,...,n$, где n – количество видов учебных работ, X_k, $k = 1,...,l$, l – количество структурных частей.

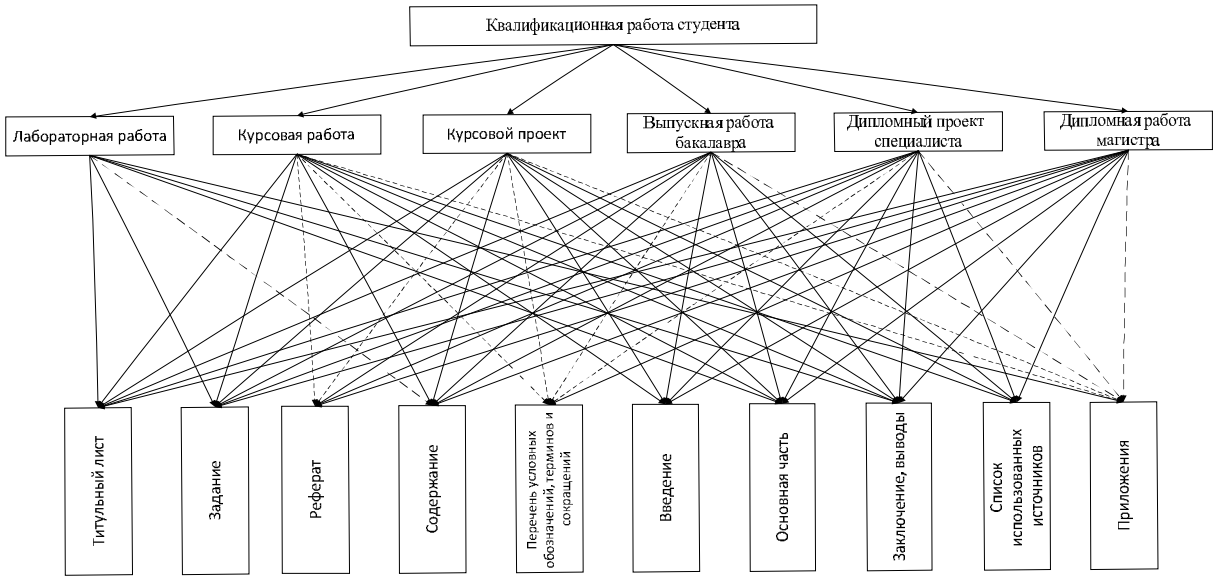

Рис. 1. Определение зависимостей между структурными частями учебных работ и уровнем важности заимствований из определенной структурной части

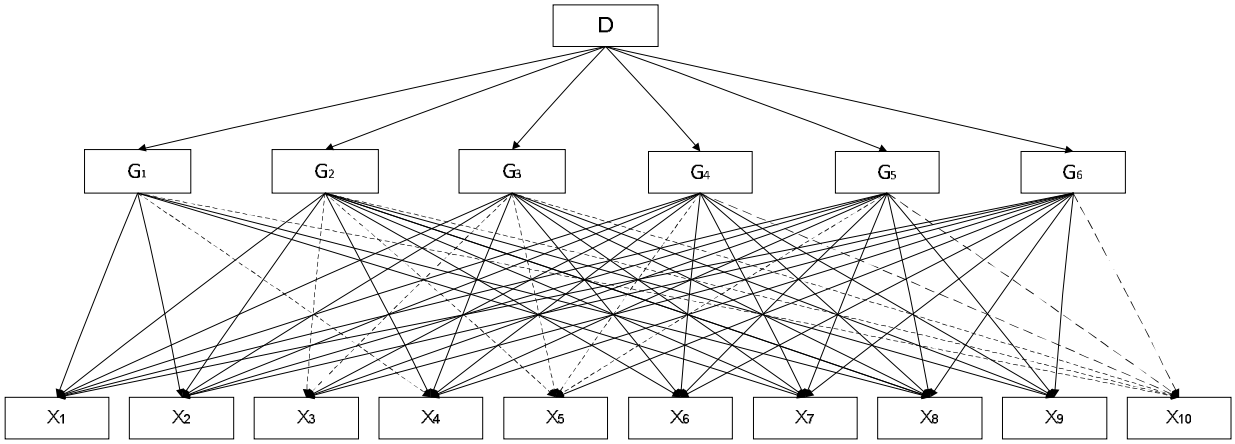

Рис. 2. Структура неформализованной задачи оптимизации

Для решения поставленной задачи необходимо доказать возможность применения метода анализа иерархий для случая, когда иерархия представляет собой совокупность нечетких множеств, т.е. необходимо доказать основную теорему метода анализа иерархий (МАИ) через теорию нечетких множеств. Применение метода анализа иерархий позволит выявить наиболее существенные структурные части (на основе ранжирования) и произвести ранжирование оценок с учетом наличия в них фактов заимствования. На рисунке 2 представлены различные упорядоченные четкие множества, которые состоят из элементов, являющимися нечеткими, и определяются каждый своей функцией принадлежности.

Элементы каждого уровня являются нечеткими подмножествами четко упорядоченного множества и определятся именно в этом смысле.

Пусть $x_1, x_2, ..., x_n$ – совокупность объектов третьего уровня иерархии. Определим веса $\omega_1, \omega_2, ..., \omega_n$ и их влияние на следующий (более верхний) уровень. Количество суждений о парах объектов (x_{n-1}, x_n) представляется матрицей размером $n \times n$. Обозначим через a_{ij} число, соответствующие значимости элемента x_n, по сравнению с x_{n-1}, а матрицу, состоящую из этих чисел – $A = \{a_{ij}\}$.

Все элементы $x \in X$ сравниваются по их соответствию понятию, представленному нечетким множеством B'. По матрице попарных сравнений A определим функцию принадлежности $\mu_{B'}(x)$ для каждого элемента $x \in X$, и элементы a_{ij} представляют собой некоторые оценки интенсивности принадлежности элементов $x_t \in X$ нечеткому множеству B' по сравнению с элементами $x_l \in X$.

Значение функции принадлежности $\mu_{B'}(x)$ для всех элементов $x \in X$ будет иметь вид

$$\mu_{B'}(x_i) = \omega_i (i \subset I \subset \{1, 2, ..., n\}). \qquad (3)$$

где $\subset$ означает строгое включение.

Определим значение уровня важности каждой из структурных частей для определенного вида квалификационной работы, для этого необходимо определить все множества элементов каждого уровня и иерархию важности данных элементов.

Набор D – полная иерархия с h уровнями. Пусть B_k – матрица приоритетов k-го уровня, $k = 2, ..., h$. Для проведения анализа учебных работ и структурных уровней иерархии примем следующие обозначения: любой k-й уровень иерархии является совокупностью отдельных элементов l_k^j, где $j = 1, ..., m$, $m \geq 1$ количество элементов данного уровня. Представим формально множество E элементов иерархии $E = \{l_k^j\}$, состоящее из подмножеств, т.е. уровней $L_к = \{l_k^j\}$. Поскольку иерархия является совокупностью уровней S_k $k = 1, ..., n$, где n - количество уровней иерархии, и отношение между уровнями $R : S_{k+1} \circ S_k \to N$ (символ $\circ$ - композиция отношений), очевидно, что каждый элемент $\{l_k^j\}$ можно считать подмножеством E

$$M_k^j = \{l_k^j\}, \forall k, j, \qquad (4)$$

при этом

$$E = \bigcup_{к=1}^{n} L_k = \bigcup_{к=1}^{n} \bigcup_{j=1}^{m} M_k^j. \qquad (5)$$

где $\cup$ –алгебраическая сумма.

Представим иерархию структурных элементов учебных работ в виде совокупности нечетких множеств, рис. 3.

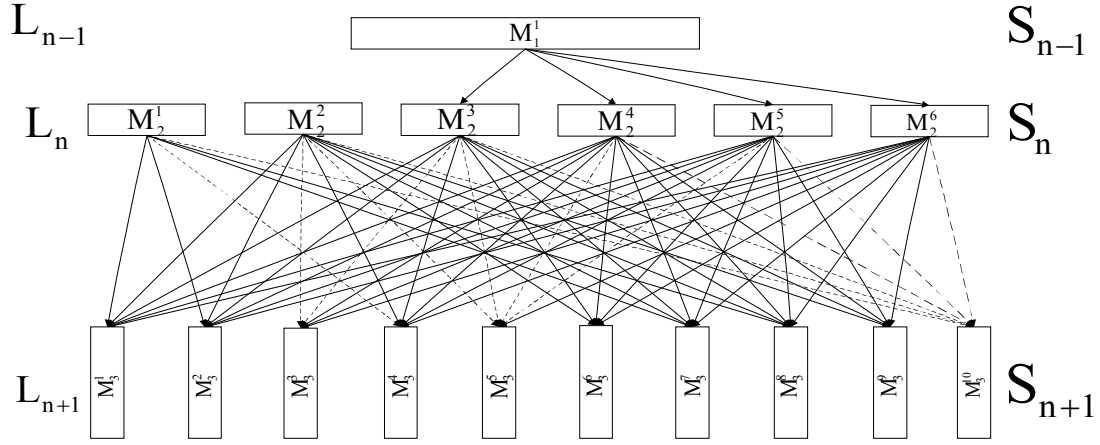

Рис. 3. Иерархическая структура учебных работ, состоящая из элементов, которые являются нечеткими множествами

Набор учебных работ M_k^j является набором нечетких свойств для элементов нижнего уровня, т.е. M_k^j - это набор нечетких множеств, при этом универсальными множествами этих нечетких множеств являются множества элементов нижних уровней.

Область определения функции принадлежности нечетких множеств M_2^1 - универсальное множество элементов уровня S_{n-1}, обозначенного L_{n-1}; при этом множество элементов уровня S_n, обозначенного L_n является универсальным (базовым) для нечеткого множества M_1^1 уровня S_{n-1}, принадлежащего множеству L_{n-1}. Запишем элементы каждого уровня.

Уровень $S_{n-1} : L_{n-1} : l_{n-1}^1 l_{n-1}^2 l_{n-1}^3 ... l_{n-1}^{m_{(n-1)}}$.

Уровень $S_n : L_n : l_n^1 l_n^2 l_n^3 ... l_n^{m_n}$.

Уровень $S_{n+1}: L_{n+1}: l^1_{n+1} l^2_{n+1} l^3_{n+1} ... l^{m_{(n+1)}}_{n+1}$.

Теорема (1). D - полная иерархия с наибольшим элементом b и h уровнями. Пусть B_k матрица приоритетов k -го уровня, $k = 2,...,h$. Если W' вектор приоритетов p - го уровня относительно некоторого элемента z в $(p-1)$ -м уровне, то вектор приоритетов W q -го уровня $(p < q)$ относительно z определяется как

$$W = B_q B_{q-1} ... B_{p+1} W' \qquad (6)$$

Из указанного выше следует, что вектор приоритетов самого нижнего уровня относительно элемента b:

$$W = B_h B_{h-1} ... B_2 W'. \qquad (7)$$

Связь соседних уровней иерархии определяется матрицей собственных векторов, т.е. матрицей приоритетов B. В общем виде данная матрица имеет вид:

$$B_{n+1} = \begin{array}{c|cccc} & l^1_n & l^2_n & ... & l^{m(n)}_n \\ l^1_{n+1} & \omega_{11} & \omega_{12} & ... & \omega_{1m_{(n)}} \\ l^2_{n+1} & \omega_{21} & \omega_{22} & ... & \omega_{2m_{(n)}} \\ l^3_{n+1} & \omega_{31} & \omega_{32} & ... & \omega_{3m_{(n)}} \\ l^4_{n+1} & \omega_{m(n+1)1} & \omega_{m(n+1)2} & ... & \omega_{m(n+1)m_{(n)}} \end{array} \qquad (8)$$

где B_{n+1} - матрица собственных векторов уровня S_{n+1}; m_{n+1} число элементов уровня $n+1$: $L_{n+1}: l^1_{n+1} l^2_{n+1} ... l^{m_{(n+1)}T}_{n+1}$; m_n число элементов уровня $n: L_n: l^1_n l^2_n ... l^{m_n T}_n$; $\omega^1_{n,\,n+1} = \omega_{11}\omega_{21}...\omega_{lm_{(n)}}{}^T$ - собственный вектор, т.е. вектор приоритетов элементов уровня S_n, относительно критерия нижнего уровня $n+1-l^1_{n+1}$.

С учетом (8) справедливы утверждения:

1) Степень принадлежности элемента l^1_n нечеткому множеству, которое является элементом l^1_{n+1} будет равно:

$$\mu_{l^1_{n+1}}(l^1_n) = \omega_{11} = \mu_{l^1_n}(l^1_{n+1}) \qquad (9)$$

2) Функция принадлежности нечеткого множества, которое является элементом l^1_{n+1} равна:

$$\mu_{l^1_{n+1}}(L^1_n) = \left[\omega_{11}\omega_{12}\omega_{13}...\omega_{1m_n}\right]^T = W^1_{n+1,n}; \quad (10)$$

- функция принадлежности первого элемента $n+1$ - го уровня, определения на базовом множестве уровня S_n (строка матрицы).

3) Функция принадлежности нечеткого множества, которое является элементом l^1_n:

$$\mu_{l^1_n}(L^1_{n+1}) = \left[\omega_{11}\omega_{12}\omega_{13}...\omega_{1m_{(n+1)}}\right]^T = W^1_{n,n+1} \quad (11)$$

- функция принадлежности первого элемента n - го уровня, определения на базовом множестве уровня S_{n+1} (столбец), эквивалентна вектору приоритетов уровня $n+1$ по самому первому критерию n - го уровня.

Функция принадлежности в общем случае имеет вид:

$$\begin{aligned} &\mu_{l^1_{n+1}}(L^1_n) = W^j_{n+1,n},\ j = 1,...,m_{n+1} \\ &\mu_{l^1_n}(L^1_{n+1}) = W^i_{n,n+1},\ i = 1,...,m_n \end{aligned} \qquad (12)$$

Из всего вышесказанного можно сделать вывод, что матрица B_{n+1} задает бинарное нечеткое отношение между нечеткими множествами, принадлежащим уровням n и $n+1$ и имеет следующий вид:

$$R_{n+1,n}: L_{n+1} \circ L_n \to [0,1]. \qquad (13)$$

Очевидно, что задача анализа иерархий сводится к нахождению $\mu_{l_1}(L_{n+1})$, что в свою очередь является эквивалентным $W_{n+1,n}$, т.е. к нахождению функции принадлежности нечеткого множества элемента первого уровня иерархии, определенного на базовом множестве уровня $n+1$.

Будем считать, что уровень S_1 состоит из одного элемента и находится в иерархии $n+1$ уровней. Отношения между уровнями вычисляются следующим образом

$$R_{n+1,n}(L_{n+1}, L_1): L_{n+1} \circ L_n \to [0,1]. \qquad (14)$$

Вычисленные по соответствующим матрицам парных сравнений приоритеты составляющих всех уровней иерархии представлены в таблице 1.

Представим каждую из составляющих нижнего уровня иерархии учебных работ в виде лингвистической переменной P с соответствующим конечным множеством состояний

$$P = (P_I - \text{низкий}, P_{II} - \text{средний}, P_{III} - \text{высокий})$$

Функция принадлежности рассматриваемой иерархии к конкретному множеству описывается следующими соотношениями:

$$\mu_{P_I}(M) = \begin{cases} 1 - 2M, & 0 \le M \le 1/2 \\ 0, & 1/2 \le M \le 1 \end{cases} \qquad (15)$$

$$\mu_{P_{II}}(M) = \begin{cases} 2M, & 0 \le M \le 1/2 \\ 2 - 2M, & 1/2 \le M \le 1 \end{cases} \qquad (16)$$

$$\mu_{P_{III}}(M) = \begin{cases} 0, & 0 \le M \le 1/2 \\ 2M - 1, & 1/2 \le M \le 1 \end{cases} \qquad (17)$$

Так как набор учебных работ M^j_k является набором нечетких свойств, то для согласования приоритетов различных уровней для учебных работ необходимо согласование данных параметров. Результат согласования представлен в табл. 1.

После проведенных расчетов оценки уровней иерархии структуры учебных работ, была построена матрица собственных векторов (табл. 1):

Таблица 1

M_3^{10} Матрица приоритетов различных уровней для учебных работ

Компоненты 1-го уровня иерархии	Приоритет 1-го уровня	Компоненты 2-го уровня иерархии	Приоритеты составляющих 2-го уровня иерархии	Компоненты 3-го уровня иерархии	Приоритеты составляющих 3-го уровня иерархии	Значение функции принадлежности			Компоненты 2-го уровня иерархии	Приоритеты составляющих 2-го уровня иерархии	Компоненты 3-го уровня иерархии	Приоритеты составляющих 3-го уровня иерархии	Значение функции принадлежности		
						P_I	P_{II}	P_{III}					P_I	P_{II}	P_{III}
M_1^1	1	M_2^1	0,08	M_3^1	0,05	1	0	0	M_2^4	0,20	M_3^1	0,02	1	0	0
				M_3^2	0,05	0,8	0,2	0			M_3^2	0,05	0	0,8	0,2
				M_3^4	0,05	0,6	0,4	0			M_3^3	0,20	0	0,6	0,4
				M_3^7	0,35	0	0	1			M_3^4	0,02	1	0	0
				M_3^8	0,25	0	0,4	0,6			M_3^5	0,02	1	0	0
				M_3^{10}	0,25	0	0,8	0,2			M_3^6	0,15	0	0,6	0,4
		M_2^2	0,11	M_3^1	0,02	1	0	0			M_3^7	0,26	0	0,1	0,9
				M_3^2	0,10	0	0,6	0,4			M_3^8	0,15	0	0,8	0,2
				M_3^3	0,05	1	0	0			M_3^9	0,03	1	0	0
				M_3^4	0,10	0	0,6	0,4			M_3^{10}	0,10	1	0	0
				M_3^5	0,03	1	0	0	M_2^5	0,23	M_3^1	0,02	1	0	0
				M_3^6	0,20	0	0,2	0,8			M_3^2	0,05	0	0,8	0,2
				M_3^7	0,24	0	0,1	0,9			M_3^3	0,20	0	0,6	0,4
				M_3^8	0,12	0	0,6	0,4			M_3^4	0,10	1	0	0
				M_3^9	0,02	1	0	0			M_3^5	0,02	1	0	0
				M_3^{10}	0,12	0,5	0,5	0			M_3^6	0,15	0	0,8	0,2
		M_2^3	0,11	M_3^1	0,02	1	0	0			M_3^7	0,28	0	0,1	0,9
				M_3^2	0,05	0	0,8	0,2			M_3^8	0,10	0	0,8	0,2
				M_3^3	0,02	1	0	0			M_3^9	0,03	1	0	0
				M_3^4	0,05	0	0,6	0,4			M_3^{10}	0,05	0	0,6	0,4
				M_3^5	0,02	1	0	0	M_2^6	0,27	M_3^1	0,02	1	0	0
				M_3^6	0,20	0	0,2	0,8			M_3^2	0,05	0	0,8	0,2
				M_3^7	0,29	0	0,1	0,9			M_3^3	0,10	0	0,6	0,4
				M_3^8	0,20	0	0,2	0,8			M_3^4	0,15	1	0	0
				M_3^9	0,03	0,8	0,2	0			M_3^5	0,02	1	0	0
				M_3^{10}	0,12	0	0,6	0,4			M_3^6	0,15	0	0,8	0,2
											M_3^7	0,23	0	0,1	0,9
											M_3^8	0,13	0	0,8	0,2
											M_3^9	0,05	1	0	0
											M_3^{10}	0,10	0	0,6	0,4

$$B_{n+1} = \begin{pmatrix} 0.12 & 0.33 & 0.55 \\ 0.18 & 0.28 & 0.36 \\ 0.084 & 0.257 & 0.659 \\ 0.19 & 0.396 & 0.414 \\ 0.17 & 0.688 & 0.592 \\ 0.24 & 0.407 & 0.353 \end{pmatrix}$$

Фокус иерархий для рассматриваемой иерархической структуры учебных работ будет равен следующему вектору значений: $(0.181 \quad 0.433 \quad 0.47)$.

Так как матрица B_{n+1} и фокус иерархий задает бинарное нечеткое отношение между нечеткими множествами, принадлежащим соответствующим уровням, то для получения четкого значения необходимо перейти от нечеткого отношения к четкому. Для этого применим центроидный метод [2, 3].Функции принадлежности будут равны

$$J(P_I) = \frac{25}{56}, \quad J(P_{II}) = \frac{34}{56}, \quad J(P_{III}) = \frac{31}{56}.$$

Используя центроидный метод, выполним четкую оценку степени важности структурных частей учебных работ по следующей формуле:

$$J^* = \sum_{i=1}^{k} J(P_L) \cdot M(L) \Big/ \sum_{i=1}^{k} M(L) \quad (18)$$

Получим:

$$J^* = \frac{\frac{25}{56}\cdot 0{,}181 + \frac{34}{56}\cdot 0{,}433 + \frac{31}{56}\cdot 0{,}47}{0{,}181 + 0{,}433 + 0{,}47} = 0{,}604$$

Из всего вышесказанного можно сделать вывод, что структурные части работ играют существенную роль при выявлении плагиата, позволяют выполнить оценку учебной работы с учетом специфики. Следует отметить, что функция принадлежности структурных частей учебных работ позволяет снизить степень субъективности и, следовательно, неопределенности при оценивании учебной работы в аспекте наличия в ней плагиата.

3. Оценка расстояний между уровнем иерархии учебных работ

Расстояния (r) между иерархиями по весовым индексам входящих в элементов нечетких множеств. Для оценки расстояний между уровнем l_1 первой иерархии, уровнем l_2 второй и l_3 третьего уровня воспользуемся описанным выше методом.

С учетом вышесказанного справедливы следующие высказывания:

1. Все элементы матрицы B положительны, т.е. $\omega_{ij} = \omega_{l_n^i} \big/ \omega_{l_{n+1}^j}$ для всех номеров $i,j = 1,2,...,m$

2. Матрица B обратно симметрична, т.е. ее элементы, расположенные симметрично относительно главной диагонали, являются обратными по отношению друг к другу. В частности, все элементы, расположенные на главной диагонали, равны единице.

3. Число структурных элементов работы одного уровня не всегда одинаково для различных видов учебных работ, т. е. если $M_1^1 \neq M_1^2$. Для устранения этой проблемы путем добавления к уровню с меньшим числом элементов нижнего уровня недостающего числа $f = |M_1^1 - M_1^2|$ элементов с нулевым весом $\nu = 0$.

После этого найдем наиболее похожие пары вершин сравниваемых уровней и частные редакционные расстояния между этими вершинами суммируются в накопитель редакционного расстояния между рассматриваемыми уровнями:

$$c(l_1, l_2) = \sum_{1}^{L_{n+1}} \left| M_1^1 - M_1^2 \right| . \qquad (19)$$

Сформируем матрицу редакционных расстояний между уровнями размером $l_1 \times l_2 \times l_3$ и ищем на ней оптимальный путь Q перевода одной иерархии в другую. Для этого необходимо направленный последовательный перебор вариантов, который обязательно приводит к глобальному максимуму. Необходимо вначале определить функцию Беллмана и оптимальные управления для всех возможных состояний на каждом шаге, начиная с последнего в соответствии с алгоритмом обратной прогонки. При этом следует учесть, то, что на последнем, n-м шаге оптимальное управление - x_n^*, определяется функцией Беллмана, которая имеет следующий вид:

$$f(S) = \max\{W_n(S, x_n)\}, \qquad (21)$$

в соответствии с которой максимум выбирается из всех возможных значений x_n, причем $x_n \in X$.

Дальнейшие вычисления производятся согласно рекуррентному соотношению, связывающему функцию Беллмана на каждом шаге с этой же функцией, но вычисленной на предыдущем шаге, с учетом вышесказанного формула (19) примет вид:

$$f_n(S) = \max\{W_n(S, x_n) + f_{n+1}(S_n(S, x_{n+1}))\} \} x_{n+1} \in X$$

Найдем величину редакционного расстояния $r(Q)$ (в нашем примере редакционное расстояние будет равно $r(Q) = 260$). Наибольшая величина расстояния R была найдена при сравнении заданных иерархий M_1^1, M_2^i и M_3^j с наиболее на них не похожей пустой иерархией. При анализе легко увидеть, что расстояние от иерархии M_3^j до любой иерархии M_i^j с числом уровней l_n принимает значение $r(i,3) = 100 + 2 \cdot 100 \cdot (l_n - 1)$. В примере $r(1,3) = 1100$, а $r(2,3) = 1900$, так что редакционное расстояние между уровнями по насыщенности таксонов находится как $r = r(Q) / r(i,3) = 260 / 1900 = 0{,}137$.

Общее редакционное расстояние $r_{общ}$ между двумя иерархиями примем равной средней величине расстояний d и r, на основе вышесказанного будет вычисляется по формуле $r_{общ} = (d + r)/2$. В нашем случае $r_{общ} = (0{,}236 + 0{,}137)/2 = 0{,}187$.

Выводы

1. Построена иерархическая модель учебной работы с учетом нечетких связей между ее структурными элементами.

2. Обосновано наличие нечеткости при выделении структурных частей учебных работ. Оценены структурные элементы учебных работ в аспекте важности данных структурных частей.

3. Получены аналитические выражения, обеспечивающие проведение оценки важности структурных элементов для различных видов работ обучающихся.

Новизна научных результатов состоит в том, что впервые учебная работа представлена в виде иерархической модели и нечетких связей между ее структурными частями; эта модель в отличие от стандартных моделей анализа иерархий, дает возможность оценить важность структурных элементов учебных работ с учетом специфики, присущей данному виду работ и неопределенностей, возникающих при соотнесении данной учебной работы с существующей номенклатурой работ.

Список литературы

1. Груздо І.В. Проблемы машинного анализа естественно-языкового текста в среде обнаружения плагиата [Текст] / І.В. Груздо // Міжнародна НТК «Інтегровані комп'ютерні технології в машинобудуванні» 2009. - Тези доповіді Том 3, – Х.: НАУ «ХАІ», 2009. – С. 60.

2. Саати Т. Принятие решений. Метод анализа иерархий: Пер. с англ. / Т. Саати. – М.: Мир, 1993. – 344 с.

3. Андерсон Т. Введение в многомерный статистический анализ / Т. Андерсон. – М.: Физматгиз, 1963. – 500 с.

Поступила в редколлегию 21.01.2013

Рецензент: д-р техн. наук, проф. И.В. Шостак, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

МОДЕЛЬ РУБРІКОВАНОГО ОБ'ЄКТА В ЗАДАЧАХ МАШИННОГО АНАЛІЗУ ТЕКСТІВ, ЩО ВРАХОВУЄ ЗНАЧУЩІСТЬ СТРУКТУРНИХ ЧАСТИН

І.В. Груздо

Розглянута формальна постановка задачі машинного аналізу природно мовних об'єктів з метою виявлення запозичень на прикладі виявлення плагіату в письмових роботах учнів. Описана нечітка структурна модель об'єкта текстологічного аналізу. Викладено методику визначення відстаней між рівнями ієрархій зазначеної моделі.

Ключові слова: *текстологічний аналіз, письмова робота, ієрархічна нечітка модель структури тексту, рівні ієрархії.*

RUBRIC MODEL IN THE PROBLEM MACHINE TEXTUAL ANALYSIS, TAKING INTO ACCOUNT THE IMPORTANCE OF STRUCTURAL PARTS

I.V. Gruzdo

Considered a formal statement of the problem of machine analysis of natural language facilities to identify borrowings for example detect plagiarism in written work students. Described fuzzy structural model object textual analysis. The technique of determining distances between levels of the hierarchy the model.

Keywords: textual analysis, written work, hierarchical fuzzy model structure of a text hierarchy.