А.В.Маликов, **А.С.Целиковский**, **Д.К.Пархоменко** (Ставрополь, СевКавГТУ). **Разработка математической модели оценки соответствия полнотекстовых документов заданному содержанию.**

В последнее время все большую актуальность в науке и практике принимают задачи эффективного поиска и проверки релевантности документов необходимым запросам. Существует частная проблема проверки соответствия учебно-методических комплексов, разрабатываемых в ОУ ВПО, требованиям ГОС ВПО. На текущий момент задача решается стандартными средствами полнотекстового поиска MS SQL Server, что не дает достаточного уровня точности и полноты отбора релевантных стандарту документов в связи с использованием традиционной модели «bag of words», не учитывающей совместное распределение слов, терминологичности, лексико-семантических особенностей языка.

В работе, представленной данным сообщением, рассматривается математическая модель, учитывающая перечисленные выше свойства. В ней: $T = \{t\}$ — множество слов; $M = \{m\}$ — существующие наборы морфологических признаков, представляются кортежами: $m = \{\mu_1, \mu_2, \ldots, \mu_n\}$, где μ_i — значение соответствующего морфологического признака в наборе, i — номер морфологического признака ($i = 1$ — часть речи, $i = 2$ — род и т. д.), $\mu_i = 0$, если данный морфологический признак в данном наборе не имеет смысла (например, род у глагола настоящего времени); $Gram(M \times M) \to \mathbf{R}$ — функция близости грамматических форм, задается в табличной форме на основе различий соответствующих морфологических признаков; $G \subset T \times M$, $g \in G$ — соответствие слов своим грамматическим формам; deriv $\subset G^2$ — отношение на G, включающее однокоренные слова; $gram$: $(G \times G) \to \mathbf{R}$; $gram(g_1, g_2) = Gram(m_1, m_2)$ — функция близости грамматических форм термов (если $g_{1,i}$ и $g_{2,i}$ не состоят в отношении derive, то $gram(g_{1,i}, g_{2,i}) = 0$), m_1, m_2 — соответствующие грамматические признаки; $T_d \subset B(G)$ — обертка для элементов $g \in G$ на случай многословных терминов и устойчивых словосочетаний, определяемых статистикой совместного использования [2]; $B(G)$ — множество подмножеств множества T; is: $(T_d \times B(G)) \to \mathbf{R}$; $is(t_d, b(G)) = \prod_{i=1}^{N_d} [NUM(t_{di}, b(G))^{-1} \sum_j^{N_b} gram(t_{di}, t_{bj})$ есть функция наличия в тексте словосочетания; $COL = \{\langle L = \{t_d | t_d \in T_d\}, W = \{w_{i,j}\}\rangle\}$ есть множество словосочетаний (единиц ГОС ВПО), представленных в виде графов (лесов) с вершинами L и матрицей лексико-семантической смежности W. Словосочетания выявляются специальным парсером на основе морфологических шаблонов и статистики (удаление стоп-слов и малозначимых слов).

Корни деревьев L являются непосредственными участниками словосочетания, а их потомки — синонимами с некоторой степенью $w_{i,j} \in W$. При этом $w_{i,i} = w_i \neq 0$, если терм t_{di} является непосредственным участником выражения с весом w_i; $w_{i,i}$ определяется на основе $idf(t_d)$ (обратной частоты встречаемости) и терминологичности терма.

Парсер на основе синтаксических шаблонов (конструкции с тире, двоеточием и т.д.) строит связи онтологии (отношения гипонимии/гиперонимии, меронимии). Вся информация помещается в граф словосочетания L:

$$\mathrm{FCOL} : COL \times B(G) \to \mathbf{R},$$

$$fcol(col_1, b(G)) = \sum_{i=1}^{n} w_{i,i} \left(1 - \prod_{j=1}^{n} (1 - w_{i,j} is(t_{dj}, b(G))\right) \Big/ \sum_{i=1}^{n} w_{i,i}, \qquad (1),$$

$t_{dj} \in col_1$ — термин, принадлежащий первому словосочетанию.

Семантический образ элемента стандарта (главы, подглавы): $S = \langle \{col\} \in \cup COL; c(col_i); w(col_i)\rangle$, где $w(col_i) = \max_{i=1}^{n}(w_{i,i})$.

Документ изначально представляется в виде $D = \langle \{g\} \in \cup G; c(g_i); w(g_i)\rangle$, где $c(g_i)$ — вектор положений термов в документе; $w(g_i) = tf(g_i)\mathrm{idf}\,(g_i)$ — вес терма в документе.

Для получения образа документа по отношению к какому-либо элементу стандарта (глава, подглава) для каждого словосочетания (запроса) используем скользящее окно размером $2n$ (n — длина словосочетания). В каждом отпечатке окна мы проверяем наличие запроса посредством (1) (запрос считается присутствующим, если $fcol(col_{\text{standard}}, b(G)_{\text{window}}) > K_{col}$).

В итоге получаем следующую структуру: $D(s) = \langle \{col\} \in \cup COL; M_{fcol}(s) | s \in S \rangle$, где $M_{fcol}(s)$ — матрица смежности найденных словосочетаний по отношению к словосочетаниям элемента s стандарта. В качестве столбцов выступают словосочетания элемента s, строк — распознанное множество $\{col\}$. Элементами матрицы являются значения $fcol$ двух словосочетаний, пересечением которых является данный элемент. Тогда степень соответствия документа элементу стандарта

$$fcol_D(s, D) = \frac{\sum_{i=1}^{n} w(col_i)_{i,i}(1 - \prod_{j=1}^{n}(1 - m_{i,j}))}{\sum_{i=1}^{n} w(col_i)}.$$

В результате выполнения работы построена модель, позволяющая учитывать совместное появление слов, терминологичность, лексико-семантические особенности языка. Настройка параметров модели (веса лексико-семантических связей, размер окна и т. д.) является отдельной исследовательской задачей.

Работа выполнена в рамках реализации ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2012 годы.

СПИСОК ЛИТЕРАТУРЫ

1. *Целиковский А. С.* Обзор методов релевантности документов запросу на примере проверки учебно-методических комплексов требованиям ГОС ВПО. — Третья международная научно-техническая конференция «ИНФОКОМ 3».
2. *Frantzi K. T.*, *Ananiadou S.*, *Jun-ichi Tsujii.* The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms. —In: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries.
3. *Браславский П.*, *Соколов Е.* Сравнение пяти методов извлечения терминов произвольной длины. — Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог».