УДК 004.852

БЛОЧНО-МОДУЛЬНАЯ СТРУКТУРА ИНФОРМАЦИОННО-ТЕРМИНОЛОГИЧЕСКОГО БАЗИСА

Серединн А.И.
Научный руководитель – д.т.н. Ковалев И.В.

Сибирский федеральный университет

При изучении любой области знаний, всегда стоит задача запоминания больших объемов информации, характерных для данной области. К ним могут относиться термины, понятия, а также взаимосвязи между ними. При этом объем применяемых терминов ограничивается предметной областью, выбранной для изучения специалистом.

Для выделения из интересующих специалиста текстов необходимых знаний следует провести предварительную обработку этих текстов. А именно, следует составить частотный словарь понятий и терминов. Наличие такого словаря позволяет при обучении учитывать статистические закономерности текстов. Естественно, что обучаемый должен быстрее и лучше усвоить те знания, которые в его текстах чаще всего встречается, т.е. следует учитывать частотные свойства текстов.

Целью применения адаптивно-обучающей технологии изучения какой-либо предметной области является интенсивное накопление специализированного словарного запаса студентов и специалистов, изучающих данную предметную область для своих профессиональных целей. Основными компонентами средств ее поддержки являются электронные частотные словари, и компьютерные системы, реализующие алгоритм обучения терминологии.

Информационно-терминологический базис строится на основе результатов анализа языкового материала. Под языковым материалом здесь следует понимать некоторое множество текстов интересующей разработчика предметной области. Размер языкового материала может варьироваться в зависимости от средств анализа, наличия оригинальных текстов и необходимого количества терминов.

Основу построения информационно-терминологического базиса составляют частотные словари.

Частотный словарь – это информация, порции которой выдаются ученику. По их данным определяется приоритетность изучения тех или иных терминов. Обучаемый должен быстрее и лучше усвоить те понятия и термины, которые чаще всего встречаются в текстах узкой предметной области, т.е. необходимо учитывать частотные свойства текстов, при этом берутся тексты из предметной области, требуемой для изучения. Использование таких словарей, полученных путем анализа языкового материала качественно улучшает процесс обучения.

Рассмотрим упрощенный алгоритм формирования информационно-терминологического базиса. Вначале необходимо выбрать тексты, на основе которых будет строиться частотный словарь. После этого задается количество понятий, которое будет изучаться. Далее строится база данных, в которой содержится ряд полей. Первое - это относительная частота встречаемости понятия в тексте (то есть, число встречаемости данного понятия в тексте, поделенное на общее количество понятий). Второе поле - это само понятие.

Также при построении частотного словаря используется и взаимосвязь понятий в тексте. В первую очередь изучаются понятия, которые находятся рядом в тексте, и,

следовательно, расположены в одном блоке. Для этого строится матрица A размерностью m×m, где m - количество понятий в тексте. Элемент матрицы a_{ij}- это количество словосочетаний понятий i и j в тексте. Тогда алгоритм формирования базиса будет следующим:

Шаг 1. Найти понятие k с наибольшей частотой.

Шаг 2. Занести понятие k в словарь.

Шаг 3. По матрице A находим наиболее часто встречающееся понятие рядом с понятием k. Назовем его h.

Шаг 4. Занести понятие h в словарь.

Шаг 5. Если понятий в словаре достаточно, то перейти на шаг 7, иначе перейти на Шаг 6.

Шаг 6. Если понятие k=понятие h, то перейти к шагу 3.

Шаг 7. Закончить процедуру.

Сформированный частотный словарь необходимо разбить на учебные порции для наиболее эффективного усвоения материала.

Рассмотрим задачу разбивки некоторого информационно-терминологического базиса с общим объемом материала Θ часов на n модулей, каждый из которых имеет объем Θ_i часов, так что

$$\Theta = \sum_{i=1}^{n} \Theta_i. \qquad (1)$$

Трудоемкость изучения каждого модуля составит:

$$R_i = e^{\lambda_i \Theta_i}(k_i \Theta_i + m_i) \ , \qquad (2)$$

Здесь:

k – коэффициент, характеризующий долю затрат на проведение различных мероприятий (получение справок и консультаций, доля затрат на выполнение контрольных мероприятий, зависящая от объема базиса);

m (часов) – трудоемкость работы по выполнению контрольных мероприятий и не зависящая от размеров базиса;

λ (1/час) - константа, показывающая скорость снижения вероятности успешного завершения изучения информационно-терминологического базиса в зависимости от его объема.

Тогда общая трудоемкость изучения информационно-терминологического базиса

$$R = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} e^{\lambda_i \Theta_i}(k_i \Theta_i + m_i) \qquad (3)$$

Выбирая количество модулей n и их объем Θ_i, можно добиться наименьшей общей трудоемкости изучения базиса.

Математически эта задача формулируется следующим образом: задан критерий (3) при условиях (1). Требуется найти такие n и $\{\Theta_1,\ldots,\Theta_n\}$, чтобы обеспечить оптимальные значения критерия (3):

$$\hat{R} = \min_{n,\Theta_1,\ldots,\Theta_n} R \qquad (4)$$

Задача (4) представляет собой задачу оптимизации нелинейного критерия при ограничениях на переменные. В аналитическом виде эта задача решается достаточно сложно, а численно может быть решена путем перебора вариантов разбивки базиса на модули, если дополнительно задать процедуру формирования таких вариантов.

Наличие экспоненциальных множителей в выражении (3) свидетельствует о том, что наилучших результатов следует ожидать, если величины λ_i и Θ_i будут равны между собой.

Поэтому имеет смысл рассмотреть задачу о разбивке базиса на модули, равновеликие по объему изучаемого материала.

Положим, что $\Theta_i=\Theta/n$, i=1,…,n, т.е. что все модули имеют одинаковый объем, кроме того, они все имеют одинаковые параметры $\lambda_i=\lambda$, $k_i=k$, $m_i=m$. Тогда критерий (3) примет вид

$$R = n \cdot e^{\lambda\Theta/n}(k\,\Theta/n + m) = e^{\lambda\Theta/n}(k\Theta + mn) \qquad (5)$$

и единственным параметром оптимизации становится n.

Для получения оптимального значения n необходимо решить относительно n уравнение $\frac{dR}{dn}=0$ или, с учетом (5), $\frac{dR}{dn} = -\frac{\lambda\Theta}{n^2}e^{\lambda\Theta/n}(k\cdot\Theta + n\cdot m) + e^{\lambda\Theta/n}\cdot m = 0.$

Отсюда получаем квадратное уравнение относительно n:

$$n^2 - \lambda\Theta n - \frac{\lambda\Theta^2 k}{m} = 0. \qquad (6)$$

Оптимальное значение $n = \hat{n}$ определяется положительным корнем уравнения (6):

$$\hat{n} = \Theta\left(\frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{m}k}\right). \qquad (7)$$

Величина $\hat{\Theta} = \frac{\Theta}{\hat{n}} = \left(\frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{n}k}\right)^{-1}$ определяет оптимальный объем модуля.

Таким образом, можно получить информационно-терминологический базис, разбитый на блоки и модули. То есть, так называемую блочно-модульную структуру базиса. В дальнейшем с помощью специальных моделей эту структуру можно проанализировать и на основе полученных данных провести ее оптимизацию по различным параметрам. А построение качественного информационно-терминологического базиса позволит повысить эффективность обучающей системы и увеличить скорость обучения на ней.