УДК 004.81

ПОСТРОЕНИЕ ПРЕДМЕТНОЙ МОДЕЛИ ОБУЧАЕМОГО НА ОСНОВЕ УЧЕБНОГО ТЕКСТА МЕТОДОМ ОСТРОВНОЙ КЛАСТЕРИЗАЦИИ

Соколова Н.А.
Херсонский национальный технический университет,
Щеголькова В.А.
Шосткинский институт Сумского государственного университета

Наиболее известным способом представления знаний об обучаемом является оверлейная модель, структура которой строится на основе предметной области. Такую модель называют предметной моделью обучаемого. Построение модели обычно выполняется вручную, т.к. она не является универсальной. Целью данной статьи является описание метода автоматизированного построения предметной модели обучаемого, используя размеченный учебный текст и ключевые слова, а также выделение логически связанных фрагментов предметной области с целью индивидуализации изложения материала, а также для тематического оценивания. Для выявления фрагментов применяется идея островной кластеризации, которая используется для тематической группировки больших корпусов текстов и сведения их в иерархическую структуру. Суть метода состоит в том, что сначала находятся области, называемые центрами островов, а затем на основании выбранной меры близости к центрам добавляются другие области, образуя тематический фрагмент.
Ключевые слова: оверлейная модель обучаемого, предметная модель обучаемого, островная кластеризация, автоматизированные системы обучения.

Введение. На сегодняшний день наиболее известным способом представления знаний об обучаемом является оверлейная модель. Она основывается на структуре предметной области, которая разбивается на элементарные единицы знаний – концепты, и хранит числовые характеристики, являясь, по сути, «функцией усвояемости». Построение оверлейной модели на практике осуществляется вручную. Зависимость от учебного материала не позволяет сделать ее универсальной.

Целью данной статьи является описание метода автоматизированного построения структуры предметной модели обучаемого (оверлейной модели) и выявление на ней логически связанных фрагментов.

В основе большинства курсов лежит учебный текст. Можно воспользоваться результатами исследований в области автоматического структурирования массивов текста. Однако обзор литературы [1, 2, 3] показал, что в общем случае эта проблема пока не имеет однозначного и эффективного решения. Существует лингвистический и статистический подходы. Первый является довольно трудоемким и выдает «зашумленные» результаты, второй не учитывает синтаксические отношения и может выдавать слишком грубое приближение к действительности.

Решение нашей задачи по сравнению с общей облегчено тем, что для обучения обычно предоставляется логически упорядоченный размеченный тематический текст, т.е. определено ограниченное количество связей и функциональное назначение отдельных фрагментов. Эти преимущества позволяют воспользоваться идеей островной кластеризации, предложенной в работе [3]. Модифицируем метод по нашу задачу.

Формализация задачи. Рассмотрим предметный материал, разделенный последовательно на небольшие области, имеющие самостоятельное функциональное значение. В начале каждой области обозначено название рубрики из предопределенного множества V_f. Например, «определение», «теорема», «доказательство», «пример», «алгоритм» и т.д. Для технических текстов такая рубрикация вполне естественна. Пусть задана нормативная модель обучаемого в виде множества $tr_j, (j = \overline{1..n})$ ключевых слов, соответствующих требованиям к знаниям и умениям.

Концептом K назовем неделимую область предметного материала, предназначенную для изучения обучаемым и оцениваемую системой. Формально концепт можно представить в виде кортежа $K = (id, tr, skeys, func)$, где id – уникальный идентификатор, tr – ключевое слово, определяемое концептом (может отсутствовать), $skeys$ – список «неопределяемых» этим концептом ключевых слов, $func$ – функциональная рубрика. Некоторые концепты могут быть не обязательными для изучения или поясняющими, например, доказательство теоремы. Тогда у них может не быть определяемого ключевого слова tr. «Неопределяемые» ключевые слова концепта $skeys$ – это слова, которые определены как ключевые в других концептах, а в данном используются как изученные ранее.

Пусть учебный текст состоит из множества концептов $K_i, (i = \overline{1..m})$. Воспользуемся логической упорядоченностью текста и пронумеруем концепты в порядке следования. На основе связи по ключевым словам можно построить ориентированный граф $G(K, E)$ концептов предметной области. Пусть множество связей E состоит из единственного отношения: концепты K_p и K_q находятся в отношении $K_p \Rightarrow K_q$ («для изучения K_q необходимо знать K_p»), если ключевое слово key_p концепта K_p находится в списке ключевых слов $skeys_q$ концепта K_q. Таким образом получим упрощенную семантическую сеть связанных концептов.

Предметная модель обучаемого может быть представлена как функция на множестве связанных концептов. Предоставление индивидуальных фрагментов материала для изучения является одной из главных ее задач. При этом сформированный фрагмент должен быть тематически связан с некоторым понятием или группой понятий.

Формирование логических фрагментов. Для выявления фрагментов воспользуемся идеей островной кластеризации, предложенной в работе [3] для тематической группировки больших корпусов текстов и сведения их в иерархическую структуру. Сначала находятся области, называемые центрами островов, а затем на основании выбранной меры близости к центрам добавляются другие области, образуя тематический кластер (фрагмент).

Модифицируем метод под решаемую задачу. На первом этапе в качестве центров выберем концепты, которые определяют ключевые слова. Вычисление меры близости основано на связности и упорядоченности концептов. На втором этапе объединим малые острова с более крупными. Определим размер острова как количество составляющих его концептов. Ограничим минимальное количество концептов острова величиной T.

В приведенном ниже алгоритме используются множества: C – множество центров; Pro – множество свободных концептов (не входящих ни в один из кластеров); Inc_i – множество концептов, инцидентных i-ому концепту; Rod_j – множество концептов-«родителей» для концепта j. Также используется понятие расстояния между концептами. Эта величина вспомогательная и обоснована последовательным построением учебного текста. Расстоянием между концептами K_i и K_j будем называть величину $|i - j|$, где i, j – номера концептов.

Входные данные: $G(K, E), \{tr_j\}$ // Граф концептов и множество ключевых слов.

Выходные результаты: $\{G_{id}(K', E')\}$ // Множество подграфов-островов.

1. Назначить концепты, определяющие ключевые слова, центрами островов.
$\forall j = 1..n$

{ $G_{id} \leftarrow K_{id}$ // Назначить концепт центром острова.

$Count_{id} = 1$ // Посчитать количество концептов острова.

$C \leftarrow id$ // Занести идентификатор конепта во множество центров.

$Pr\,o \rightarrow K_{id}\}$ // Исключить концепт из множества свободных концептов.

2. Составим рекурсивную функцию формирования подграфа центра. Для этого выделим множество инцидентных центру концептов. Если концепт имеет единственного родителя в виде центра, то свяжем их. Если существуют еще и другие родители, то выберем последний по номеру, и привяжем блок к нему. Выбор последнего обусловлен тем, что материал располагается последовательно и, зачастую, чтобы понять следующую часть, нужно знать предыдущую. Привязка к последнему гарантирует, что предыдущие области материала изучены ранее.

Входные данные: Rod – центр острова

Функция *Add(Rod)*

Inc_{Rod} // Сформировать множество инцидентных центру концептов из множества свободных

Если $Inc_{Rod} \neq \varnothing$, то // Если такие имеются, то для каждого выполнить

$\{\forall K_i \in Inc_{Rod}$

Если связь $Rod \Rightarrow K_i$ единственна для K_i, то

$\{G_{idRod} \leftarrow K_i$ // Занести K_i в остров родителя Rod

$Count_{idRod} = Count_{idRod} + 1$ // Увеличить количество концептов острова

$Pr\,o \rightarrow K_i$ // Удалить K_i из множества свободных концептов

$Add(K_i)\}$ // Применить функцию Add к K_i

Если связь $Rod \Rightarrow K_i$ не единственна для K_i, то

$\{Rod_{Ki}$ // Определить множество концептов-родителей для K_i

$idmax(Rod_{Ki})$ // Найти максимальное id для этого множества

$G_{idmax} \leftarrow K_i$ // Занести K_i в остров с идентификатором $idmax$

$Count_{idmax} = Count_{idmax} + 1$ // Увеличить счетчик количества концептов

$Pr\,o \rightarrow K_i\}$ // Удалить K_i из множества свободных концептов

} // Конец функции

3. Отнесем свободные концепты к островам (вызов рекурсивной функции).

$\forall id \in C$

$\{Add(K_{id})\}$

4. Может оказаться так, что некоторые острова будут маленькими. В этом случае следует склеить их с более крупным и подходящим по смыслу островом. Ограничим минимальное количество концептов острова величиной T.

$\forall id \in C$, где $Count_{id} < T$ // Для всех островов, размеры которых меньше T

// Присоединим их к ближайшему острову среди родителей и потомков

$\{id\,min = \min_{l \in Inc_{Kid}, Rod_{Kid}} (id_l - id)$ // Вычислить номер ближайшего острова

$G_{idmin} \leftarrow G_{id}$ // Присоединить текущий остров к найденному соседу

$Count_{idmin} = Count_{idmin} + Count_{id}$ // Изменить количество концептов

$C \rightarrow id\}$ // Удалить id маленького острова из списка концептов

В результате получим множество тематических островов, каждый из которых можно считать смысловым фрагментом.

Результаты. Для иллюстрации выбран структурированный фрагмент лекции по дискретной математике «Дизъюнктивные нормальные формы», разделенный на концепты.

К 0. Определение Логическая функция.

К 1. Определение

Логическое произведение нескольких переменных, взятых с отрицанием или без него, называется **элементарным произведением** или **элементарной конъюнкцией**.

К 2. *Пример*

$x_1\overline{x_2}$, $x_1\overline{x_2}x_3$ – это *элементарная конъюнкция*, $\overline{x_1x_2}$ – не *элементарная конъюнкция*.

К 3. Определение

Логическая функция, представленная дизъюнкцией *элементарных конъюнкций*, называется **ДНФ**.

К 4. Теорема 1

Элементарное произведение равно 1 тогда и только тогда, когда переменным с отрицанием присвоен нуль, а переменным без отрицания – 1.

К 5. Д-во: Если переменным с отрицанием в *элементарной конъюнкции* присвоен 0, то.

К 6. Следствие

Каждому *элементарному произведению* соответствует один и только один набор значений, входящих в него переменных, на котором оно равно 1.

К 10. Определение

Логическая функция n переменных, принимающая значение, равное единице, только на одном их наборе, называется **конституэнтой 1**.

К 11. Правило записи *конституэнты 1.*

К 14. Определение

Дизъюнкция *конституэнт 1*, равная единице на тех же наборах, что и заданная функция, называется **совершенной дизъюнктивной нормальной формой** (**СДНФ**).

К 15. Теорема

Любая *логическая функция* F, за исключением константы нуля, единственным образом представима в *СДНФ*.

К 16. Правило построения *СДНФ* (запись *логической функции* по единицам).

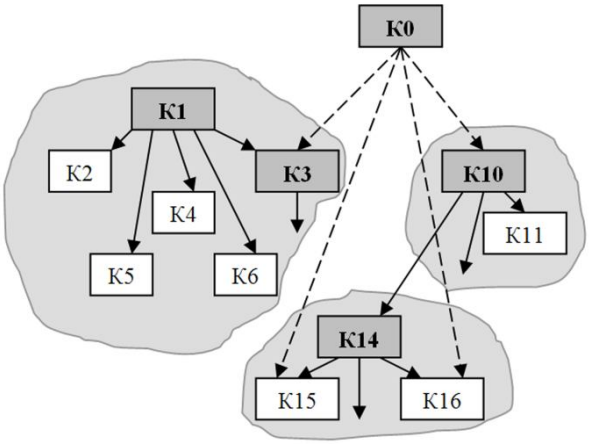

Рисунок 1 – Структура предметной модели обучаемого с выделенными тематическими фрагментами

Рубрики: «определение», «пример», «теорема», «доказательство», «следствие», «правило». Определяемые ключевые слова: элементарное произведение (элементарная конъюнкция), ДНФ, конституэнта 1, СДНФ. Курсивом показаны «неопределяемые» в текущем концепте ключевые слова, используемые для связи концептов. На рис. 1 представлен подграф предметной области и тематические фрагменты, полученные

с помощью метода островной кластеризации. К0, К1, К3, К10, К14 – центры островов. К1 и К3 образуют укрупненный остров. К0 – концепт из предыдущей лекции, являющийся обеспечивающим материалом для данной лекции.

Индивидуализацию обучения можно обеспечить за счет коррекции фрагментов на основании таких параметров модели обучаемого, как уровень знаний, список ошибок, способности к обучению и т.д. Фрагменты можно использовать при формировании уроков, на них можно определять уровень усвоения материала, а также применять для обнаружения пробелов в знаниях.

Выводы. В результате использования предложенного метода можно получить множество тематических фрагментов предметной области, которые составят основу предметной модели обучаемого. Полученная графовая структура является картой знаний обучаемого. В дальнейших исследованиях можно расширить метод для получения фрагментов, находящихся в зоне ближайшего развития обучаемого, выявления и подбора корректирующих фрагментов, ввести функцию оценивания отдельных концептов и островов. Нужно отметить, что метод хорош для технических дисциплин, однако проблематичен для гуманитарных из-за слабой структуризации учебного текста. Кроме того актуальна проблема выявления синонимов.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Виноградова Н. В. Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике / Н. В. Виноградова, О. А. Митрофанова, П. В. Паничева // Труды девятой Всероссийской научной конференции «Электронные библиотеки : Перспективные методы и технологии, электронные коллекции» (RCDL–2007). Переславль-Залесский, 2007 : [Электронный ресурс]. – Режим доступа : http://www.rcdl.ru//papers/2007/paper_31_v1.pdf

2. Ермаков А. Е. Автоматизация онтологического инжиниринга в системах извлечения знаний из текста / А. Е Ермаков //Материалы международной конференции «Диалог 2008». – М. – С. 154-159.

3. Киселев М. В. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики / М. В. Киселев, В. С. Пивоваров, М. М. Шмулевич // Интернет-математика. Автоматическая обработка веб-данных. – М., 2005. – С. 412-435.

Соколова Н.А., Щеголькова В.О. ПОБУДОВА ПРЕДМЕТНОЇ МОДЕЛІ УЧНЯ НА ОСНОВІ УЧБОВОГО ТЕКСТУ МЕТОДОМ ОСТРІВНОЇ КЛАСТЕРИЗАЦІЇ

Оверлейна модель є найбільш відомим способом представлення знань про учня в автоматизованих системах навчання. Її структура будується на основі предметної області. Таку модель називають предметною моделлю особи, що навчається. У даній статті пропонується автоматизувати процес побудови предметної моделі особи, що навчається, використовуючи розмічений учбовий текст і ключові слова. Крім того, вирішується завдання виділення логічно зв'язаних фрагментів предметної моделі з метою індивідуалізації подання матеріалу, а також для тематичного оцінювання. Для виявлення фрагментів застосовується ідея острівної кластеризації, яка використовується для тематичного групування великих корпусів текстів і зведення їх в ієрархічну структуру. Метод полягає в тому, що спочатку знаходяться області – центри островів, а потім на підставі обраної міри схожості до центрів додаються інші області, утворюючи тематичний фрагмент.

Ключові слова: оверлейная модель учня, предметна модель особи, що навчається, острівна кластеризація, автоматизовані системи навчання.

Sokolova N.A., Shegolkova V.A. STUDENT MODELING BASED ON THE TRAINING TEXT CLUSTERING METHOD ISLAND

The most famous way of representing knowledge about the student is the overlay model, the structure of which is based on the domain. Such a model is called the subject student model. Building a model is usually done by hand, as it is not universal. In this paper we propose to automate the process of constructing the subject of the student model using labeled training text and keywords. In addition, the allocation problem is solved logically related fragments subject the model to customize the presentation of the material, as well as the thematic evaluation. Fragments used to identify clustering idea island, which is used for the thematic grouping of larger corpus and information in their hierarchical structure. The method works with the subject area, broken into disjoint text area. The method consists in the fact that the first field are called centers of the islands, and then selected based on a measure of proximity to the centers of other areas are added to form a thematic fragment.

Keywords: overlay student model, object model trainee island clustering, automated learning systems.

Статтю прийнято
до редакції 11.06.14.