УДК 004.82

ВЫЯВЛЕНИЕ СТРУКТУРЫ УЧЕБНО-СПРАВОЧНЫХ МАТЕРИАЛОВ И ФОРМИРОВАНИЕ ТРАЕКТОРИЙ ИХ ОСВОЕНИЯ

Алипова Н.А.[1]

[1]*ФГБОУ ВПО «Нижегородский государственный технический университет им. Р.Е. Алексеева», Нижний Новгород, Россия (603950, ГСП-41, г. Н. Новгород, ул. Минина, д. 24), e-mail: Alipovana@mail.ru*

Повышение эффективности освоения учебно-справочных материалов возможно за счет определения порядка изучения, при котором полученных ранее знаний будет достаточно для освоения каждого следующего фрагмента контента. Для этого требуется построить структуру предметной области, соответствующей имеющемуся контенту. Построить семантическую модель на уровне прикладной задачи вручную, не имея достаточных знаний в соответствующей предметной области, весьма затруднительно. Для выявления структуры учебно-справочных материалов предлагается сначала разделить их по темам, т.е. сформировать кластеры, содержащие учебный контент сходной тематики. Затем построить траектории освоения материала, определяющие последовательность изучения на различных иерархических уровнях. Применение предложенного подхода позволяет выявить, а также представить в наглядном виде взаимосвязи фрагментов контента, их тематическую близость, построить сеть знаний фрагментов контента и траектории освоения материала.

Ключевые слова: тезаурус, контент, мера близости, кластеризация, сеть знаний, траектория освоения материала.

EDUCATIONAL ANDREFERENCE SOURCE STRUCTURE DETECTION AND LEARNING TRAJECTORY FORMATION

Aliopova N.A.[1]

[1]*Nizhny Novgorod State Technical University n.a. R.E. Alekseev, Nizhny Novgorod, Russia (603950, Nizhny Novgorod, Minin street, 24), e-mail: Alipovana@mail.ru*

Improving the educational and reference materials study efficiency is possible by determining the order of the study, in which the previously obtained knowledge volume will be sufficient for the learning of the each next content fragment. This requires building of the structure of the domain corresponding to the existing content. Build a semantic model for application level tasks manually without having sufficient knowledge of the domain, is rather difficult. To identify the structure of educational and reference materials, first proposed to divide them into topics, i.e. form clusters containing similar subjects learning content. Then build a learning trajectory, determining the sequence of study at various hierarchical levels. Application of the proposed approach allows to identify, as well as to visualize the relationship between content fragments, their thematic affinity, to build a content fragments knowledge network and learning trajectories.

Keywords: thesaurus, content, measure of adjacency, clusterization, knowledge network, learning trajectory.

Введение

Повышение эффективности изучения учебно-справочных материалов (таких, как ресурсы wiki, словари BaseGroup, а также глоссариев различных программных пакетов, например Statistica, Mathcad, Matlab и др.) возможно за счет определения порядка изучения, при котором полученных ранее знаний будет достаточно для освоения каждого следующего фрагмента контента. Для этого нужна структура предметной области, соответствующая имеющемуся контенту. Примерами моделей таких структур являются классификаторы УДК, ГРНТИ и т.п. Однако они охватывают слишком широкую предметную область, за счет чего сложны для поиска отдельного раздела. Другим недостатком является недостаточная детализация таких классификаторов – они заканчиваются практически на том уровне, с которого начинаются классификации понятий предметных областей прикладных задач.

Справочные информационные ресурсы, как правило, упорядочены по алфавиту. Построить семантическую модель на уровне прикладной задачи вручную, не имея достаточных знаний в соответствующей предметной области, весьма затруднительно.

Для выявления структуры учебно-справочных материалов предлагается сначала разделить их по темам, т.е. сформировать кластеры, содержащие учебный контент сходной тематики. Затем определяются траектории освоения материала, определяющие последовательность изучения на различных иерархических уровнях. Таким образом, для формирования структуры учебно-справочных материалов необходимо выполнить следующие этапы.

1. Индексация всего множества фрагментов учебного контента.
2. Кластеризация фрагментов учебного контента.
3. Упорядочение фрагментов контента внутри кластеров – формирование траекторий освоения материалов нижнего уровня.
4. Упорядочение полученных кластеров – формирование траектории освоения материалов верхнего уровня.

На первом этапе, с целью автоматизации процесса структурирования учебно-справочных материалов, выбранные фрагменты контента индексируются на основе тезауруса запросов обучающихся одним из способов, рассмотренных в [1]. Разработана программа, позволяющая автоматически формировать тезаурус на основе указания одной наиболее общей (родительской) темы [2].

Для примера выберем фрагменты электронного учебного контента в области информационных систем и технологий по нескольким темам и обозначим их C_n, $n=\overline{1,N}$ (в выбранном примере $N=8$): «Системы управления базами данных» (C_1– СУБД), «Базы данных» (C_2– БД), «Модель данных» (C_3– МД), «Предметная область» (C_4– ПО), «Экспертные системы» (C_5– ЭС), «Корпоративные информационные системы» (C_6– КИС), «Информационные системы» (C_7– ИС), «Информационные технологии» (C_8– ИТ).

Пусть тезаурус содержит термины, соответствующие названиям одноименных статей, и расширяется двумя дополнительными терминами: «Реляционная модель» и «MySQL», которые близки по тематике к выбранным темам (ссылки на них несколько раз встречаются в рассматриваемых фрагментах контента). Таким образом, формируется базис B, содержащий элементы B_m, $m=\overline{1,M}$, являющийся основой для определения индексов фрагментов учебного контента.

Для индексации контента строится матрица **I** (Index), строки которой являются индексами фрагментов контента, т.е. в строках содержатся коэффициенты разложения по

базисным векторам для соответствующих фрагментов контента. Вхождение элементов базиса определяется, в данном случае, по наличию ссылок из фрагмента контента на каждый из элементов базиса: если ссылка есть – 1, если нет – 0 (таблица 1), т.е. элемент матрицы $I_{ij} = 1$, если фрагмент *i* содержит ссылки на *j*-й элемент базиса, и 0 – в противном случае.

Таблица 1. Матрица I бинарных индексов фрагментов контента

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
C_1	1	1	1	0	0	0	1	1	1	1
C_2	1	1	1	1	0	0	0	0	1	1
C_3	1	1	1	0	0	0	1	0	1	1
C_4	0	0	0	1	0	0	0	0	0	0
C_5	0	1	0	1	1	0	1	0	0	0
C_6	0	1	1	0	0	1	0	1	0	0
C_7	1	1	0	0	0	1	1	1	0	0
C_8	0	0	0	0	0	0	1	1	0	0

На втором этапе для решения задачи кластеризации воспользуемся бинарным индексом фрагментов контента. При кластеризации алгоритмом k-means, в зависимости от выбора количества кластеров q (с использованием пакета DeduktorBasegroup), получены следующие результаты:

q=2: {C1, C2, C3}, {C4, C5, C6, C7, C8}; (1)

q=3: {C1, C2, C3}, {C4, C5}, {C6, C7, C8}. (2)

Обычно кластеризация рассматривается как метод обучения без учителя, который группирует объекты на основе только той информации, которая представлена в самих множествах объектов и не использует дополнительную информацию. В реальных задачах зачастую возникают дополнительные ограничения и условия, для учета которых необходима дополнительная информация об объектах и их связях или о размерах кластеров, которые желательно получить. Использование такой информации может существенно облегчить обработку результатов кластеризации (т.к. полученные результаты уже будут учитывать ограничения задачи). Однако традиционные алгоритмы, например k-means, не предоставляют механизма учета такой информации.

Альтернативой применения алгоритма k-means является кластеризация с помощью построения дендрограмм [1]. Объединяя «снизу вверх» уровни дендрограммы до тех пор, пока количество элементов в кластерах не превышает заданные ограничения, можно получить «оглавление» выбранной предметной области, т.е. распределение фрагментов учебных материалов по темам.

Исходными данными для задачи кластеризации с ограничениями на размер кластеров является количество объектов в каждом кластере. При этом, учитывая, что общее количество объектов известно, можно рассчитать минимальное количество кластеров для такого разбиения. В отдельных задачах количество кластеров может быть задано.

Пусть в рассматриваемом примере ограничение на размер кластера составляет 3 элемента, т.е. в каждый из кластеров должно входить не более трех элементов.

Тогда из представленных результатов кластеризации ограничению задачи удовлетворяет только разбиение (2), т.к. в разбиении (1) присутствует кластер с пятью элементами.

На третьем этапе формируются траектории освоения материалов нижнего уровня.

Траектории нижнего уровня, представляющие собой последовательность объектов, входящих в отдельные кластеры, могут быть построены:

1) экспертным путем, на основе определения парных дидактических связей;
2) в автоматизированном режиме, с помощью процедур анализа содержимого фрагментов контента.

В первом случае потребуется процедура проверки согласованности мнений группы экспертов, а также мнений каждого эксперта в отдельности, с целью исключения циклического упорядочения фрагментов контента. Кроме того, трудоемкость такой процедуры существенно возрастает с увеличением количества фрагментов контента. Автоматизация процесса выявления связей между фрагментами контента позволит снизить общую трудоемкость данного этапа, а также повысить точность определения весов связей.

Для автоматизированного определения связей между фрагментами контента и их весов применяется второй способ, основанный на расчете несимметричных мер включения, характеризующих степень включения одного материала в другой.

Воспользуемся несимметричным коэффициентом Жаккара [10]:

$$K(C_i; C_j) = \frac{n(C_i \cap C_j)}{2n(C_i) - n(C_i \cap C_j)}, \quad (3)$$

где C_i – вектор коэффициентов разложения по базису *i*-го фрагмента контента.

Элементы матрицы $\mathbf{S_A}$ близости фрагментов контента: $S_{A_{ij}} = K(C_i, C_j)$ (таблица 2).

Для формирования траекторий изучения материала на нижнем уровне сначала строятся сети знаний для объектов каждого отдельного кластера с использованием матрицы несимметричных мер включения. Для каждого из трех кластеров (2) построены сети знаний (рис. 1), в узлах которых расположены фрагменты контента, связи обладают весом, показывающим близость фрагментов между собой, а также направлением, характеризующим их дидактическую упорядоченность.

Таблица 2. Матрица несимметричных мер включения для фрагментов контента C_n

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
C_1	1	0.556	0.75	0	0.167	0.273	0.4	0.167
C_2	0.714	1	0.714	0.091	0.2	0.2	0.2	0
C_3	1	0.714	1	0	0.2	0.2	0.333	0.091

C_4	0	1	0	1	1	0	0	0
C_5	0.333	0.333	0.333	0.143	1	0.143	0.333	0.143
C_6	0.6	0.333	0.333	0	0.143	1	0.6	0.143
C_7	0.667	0.25	0.429	0	0.25	0.429	1	0.25
C_8	1	0	0.333	0	0.333	0.333	1	1

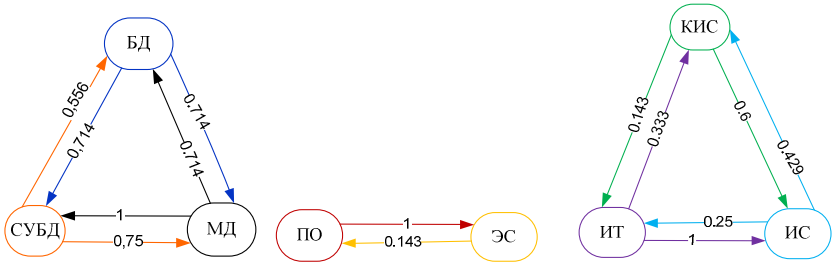


Рис. 1. Сети знаний для трех кластеров.

Как правило, топология каждой или одной из таких подсетей близка к случаю полного графа. Тогда актуальной становится задача выделения наиболее важных связей и «отбрасывания» наименее существенных. С целью выделения наиболее весомых связей строится минимальный каркас. Для этого в качестве веса выбирается расстояние относительно несимметричных мер включения, определяемое как дополнение их до единицы, и строится минимальный каркас, используя все наиболее важные связи. Затем полученный каркас рассматривается как сеть знаний, которая характеризуется отсутствием направленных циклов и транзитивных связей (рис. 2).

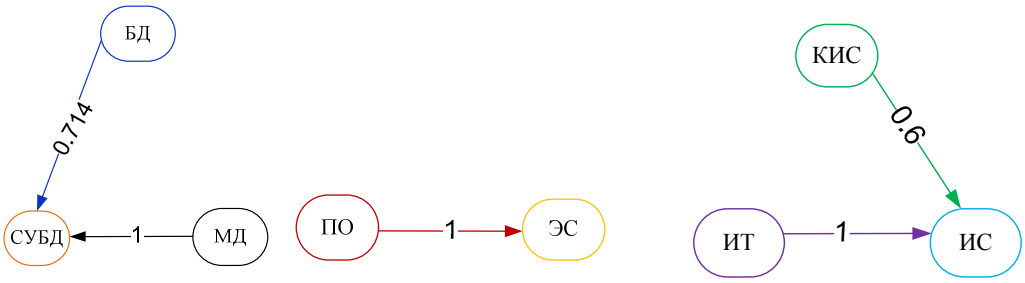


Рис. 2. Минимальные каркасы – сети знаний кластеров без циклов.

На основе обработки сети знаний предложена процедура определения последовательности освоения учебно-справочных материалов. Поскольку сеть знаний является взвешенным орграфом, то к ней применимы алгоритмы на графах. В терминах теории графов построение траектории соответствует построению *простой цепи*, т.е. такого маршрута, у которого все ребра различные и не содержится одинаковых вершин. Наиболее близким к задаче упорядочения данной сети является алгоритм топологической сортировки графа. Введем на вершинах графа частичное отношение порядка, а именно: вершина i считается меньше чем вершина j, если в графе есть ребро из i в j, т.е. есть связь между

фрагментами контента C_i и C_j. Задача топологической сортировки состоит в том, чтобы построить для данного графа такой порядок обхода его вершин, чтобы «меньшая» вершина стояла в этом порядке позже, чем «большая» [3]. Таким образом, будет получена траектория освоения материала (рис. 3). Недостаток данного алгоритма в том, что он не учитывает веса связей, а только их направления. Поэтому для формирования траектории освоения материала предварительно потребовалось построить минимальные каркасы и к ним применять топологическую сортировку таким образом, чтобы вес связей был учтен на «подготовительном» этапе.

[МД]–>[БД]–>[СУБД]

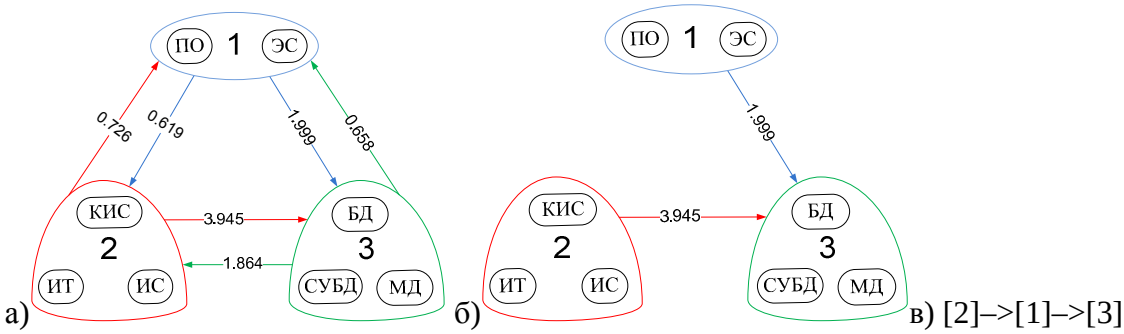
[ПО]–>[ЭС]

[ИТ]–>[КИС]–>[ИС]

Рис. 3. Траектории освоения материала нижнего уровня.

На четвертом этапе необходимо построить траектории верхнего уровня, т.е. упорядочить кластеры, полученные на втором этапе. Для этого построим сеть знаний, узлами которой будут полученные кластеры. Для определения связей между кластерами и их весов воспользуемся несимметричными коэффициентами Жаккара, рассчитанными на третьем этапе. Рассмотрим связи, которые соединяют объекты, вошедшие в разные кластеры.

Пусть множество таких связей L. Разделим его на подмножества L_{mn}, каждое из которых содержит связи, направленные от одного кластера (кластера n) к другому (кластеру m). Количество K_L таких подмножеств будет равно количеству упорядоченных пар полученных кластеров, т.е.: $K_L = \frac{k!}{(k-2)!}$, где k – количество кластеров, полученных на втором этапе. Другими словами, суммы элементов подмножеств L_{mn}, т.е. $\sum_{L_{mn}} S_{A_{ij}}$ также можно рассматривать как несимметричные характеристики сходства, на основе которых может быть построена сеть знаний с кластерами в узлах (рис. 4а). Используя вышеописанную процедуру упорядочения элементов сети знаний (т.е. построение минимального каркаса и последующей его топологической сортировки), можно получить последовательность освоения тем, т.е. траекторию верхнего уровня (рис. 4 б, в).



а) б) в) [2]–>[1]–>[3]

Рис. 4: а) сеть знаний с кластерами в узлах, б) её минимальный каркас, в) траектория верхнего уровня.

В результате автоматически сформированная последовательность освоения учебно-справочных материалов может быть представлена оглавлением:

Тема 1

1.1 Информационные технологии.

1.2 Корпоративные информационные системы.

1.3 Информационные системы.

Тема 2

2.1 Предметная область.

2.2 Экспертные системы.

Тема 3

3.1 Модель данных.

3.2 Базы данных.

3.3 Система управления базами данных.

Применение автоматизированных процедур индексации и кластеризации фрагментов учебного контента позволяет выявить, а также представить в наглядном виде взаимосвязи этих фрагментов, их тематическую близость, построить сеть знаний фрагментов контента и траектории освоения материала.

Список литературы

1. Баранов В.Г., Милов В.Р., Алипова Н.А., Егоров Ю.С. Подход к представлению материалов в информационно-обучающих системах // Информационно-измерительные и управляющие системы. – 2013. – Т. 11, № 7. – С. 19-23.
2. Баранов В.Г., Милов В.Р., Егоров Ю.С., Алипова Н.А., Курушин А.Н. Формирование структуры wiki-ресурсов : свидетельство о государственной регистрации программы для ЭВМ). № 2012619227. Заявка № 2012616892. Приоритет 15.08.2012 г. Рег. номер 2012619227 (12.10.2012). 1 с.
3. Дятлов С. Алгоритмы на графах [Электронный ресурс] // Программы для учителя : сайт. – URL: http://teasoft.ru/data/algoritm/graf.htm (дата обращения: 24.03.2014).
4. Куркин А.А., Максимов М.Ю. Дискретная математика : учебное пособие / НГТУ им. Р.Е. Алексеева. – Нижний Новгород, 2013. – 145 с.
5. Мелик-Гайказян И.В. Информационные процессы и реальность. – М. : Наука. Физматлит, 1998. – 192 с. – ISBN5-02-015086-X.

6. Милов В.Р., Баранов В.Г., Эпштейн А.Ю., Суслов Б.А. Оценка характеристик логических правил на основе байесовской методологии // Информационно-измерительные и управляющие системы. – 2011. – Т. 9, № 3. – С. 56-60.
7. Харламов А.А. Нейросетевая технология представления и обработки информации (естественное представление знаний). Кн. 4 : монография / под ред. А.И. Галушкина. - М. : Радиотехника, 2006. – 88 с. : ил. (Научная серия «Нейрокомпьютеры и их применение», редактор А.И. Галушкин). - ISBN5-88070-073-9.
8. Шрейдер Ю.А. Тезаурусы в информатике и теоретической семантике // НТИ. Сер. 2. – 1971. – N 3.
9. Нгуен Ба Нгок, Тузовский А.Ф. Классификация текстов на основе оценки семантической близости терминов // Известия Томского политехнического университета. - 2012. - Т. 320, № 5. – С. 43-48.
10. Мера сходства // Википедия – свободная энциклопедия [Электронный ресурс]. – URL: http://ru.wikipedia.org/wiki/Мера_сходства (дата обращения: 16.04.2014).

Рецензенты:

Милов В.Р., д.т.н., профессор, заведующий кафедрой «Электроника и сети», Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Нижегородский государственный технический университет им. Р.Е. Алексеева», Минобнауки России, г. Нижний Новгород.

Мисевич П.В., д.т.н., профессор, профессор кафедры «Вычислительные системы и технологии», Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Нижегородский государственный технический университет им. Р.Е. Алексеева», Минобнауки России, г. Нижний Новгород.