УДК 025.4:004.912

О.В. Бисикало

Ассоциативный поиск для задач обучения на основе электронного тезауруса образов

Статья посвящена созданию «умного» контента для электронных учебников. Рассмотрена реляционная модель тезауруса образов, наполняемого в результате обработки текстов с учебным материалом. Предложены формальные методы ассоциативного поиска типа «ИЛИ», «И» и ответ на вопрос. Действующий прототип системы программно реализован на основе технологии *Python + SQLite + Graphviz*.

The paper is devoted to the creation of the «clever» content for electronic textbooks. The relational model of the image thesaurus being filled as a result of the manipulation of texts with the teaching aids is considered. The formal methods of an associative search for the type «OR», «AND», an answer to a question are suggested. The operating prototype of the system is programmatically realized on the basis of the Python + SQLite + Graphviz technology.

Статтю присвячено створенню «розумного» контенту для електронних підручників. Розглянуто реляційну модель тезаурусу образів, що заповнюється в результаті обробки текстів з навчальним матеріалом. Запропоновано формальні методи асоціативного пошуку типу «АБО», «І» та відповідь на питання. Діючий прототип системи програмно реалізовано на основі технології *Python + SQLite + Graphviz*.

Введение. Развитие современных мультимедийных компьютерных технологий, в том числе *Semantic WEB* и виртуальных миров, направлено, в конечном итоге, на повышение эффективности восприятия человеком информации из компьютера. Тем не менее, несмотря на значительный прогресс в данной области подавляющая часть человеческих знаний все еще заключена в текстовой форме [1]. В задачах электронного обучения серьезный перевод знаний из символьно-вербального формата в мультимедийный требует материально-технических затрат, несовместимых с сегодняшним уровнем финансирования образования Украины. Один из возможных путей решения проблемы повышения эффективности *e*-обучения – разработка методов и средств извлечения семантической информации из естественно-язычных текстов [2] с целью уменьшения непродуктивного времени функционирования обучающей системы.

В работе [3] предлагается подход к построению электронного тезауруса как словаря образов предметной области, исходя из требований общей модели образного мышления. В рамках подхода простое повествовательное предложение (синтагма текста) рассматривается как аналог события, где образы объединяются между собой с помощью ассоциативных связей [2]. В результате использования предложенного в [3] редактора для ввода обучающего контента появляется возможность дополнить традиционный гипертекст дополнительными синтагматическими связями между образами. Однако достижение нового качества контента не подкреплено реализацией типов образного поиска, отсутствующих в аналогичных системах.

Постановкой задачи будем считать разработку формальных методов ассоциативного поиска на основе использования электронного тезауруса образов.

Формализация тезауруса образов

Исходя из принципов построения модели образного мышления [2] предлагается рассматривать обучающую систему как экспертную, предварительно заполняемую знаниями на основе обработки текстов учебного содержания. Если блок памяти такой системы реализован в виде реляционной базы данных [3], то часть нетривиальных задач обучения представляется путем комбинации базовых типов ассоциативного поиска, получаемых с помощью *SQL*-запросов к базе данных. Тогда тезаурус образов формализуется на основе следующих отношений:

• Обучающие тексты (учебные дозы) и составляющие их фразы (синтагмы) представлены в системе как отношения

$$Text - RE \subset Text - Id \times Bi - \\ - Te \times Title \times Author \times Time, \qquad (1)$$

где *Text-Id* – уникальный код текста, *Bi-Te* – двоичный код учебной дозы, *Title* – название текста, *Author* – автор текста, *Time* – время внесения дозы в систему и

$$Event \subset Event - Id \times Bi - \\ - Sy \times Text - Id \times Syntagma, \quad (2)$$

где *Event-Id* – уникальный код синтагмы, *Bi-Sy* – двоичный код учебной фразы, *Syntagma* – вербальное обозначение фразы.

• Собственно словарь образов представлен в виде отношения

$$Image \subset Bi - I \times Object - Quality \times Object \times \\ \times Notion \times Method \times Method - Quality, \quad (3)$$

где *Bi-I* – двоичный код образа и вербальные обозначения: *Object-Quality* – качество объекта, *Object* –объект, *Notion* – понятие, *Method* – метод, *Method-Quality* – качество метода.

• Элементарным конструктом предложения является ассоциативная пара образов, представленная в виде отношения

$$Assoc - Twice \subset Bi - I_1 \times Bi - \\ - I_2 \times Twice - Id \times Force \times Force^{-}, \quad (4)$$

где *Bi-*I_1 – двоичный код первого образа пары, *Bi-*I_2 – двоичный код второго образа пары, *Twice-Id* – уникальный код пары, *Force* – значение силы прямой синтагматической связи между образами, $Force^{-}$ – значение силы обратной синтагматической связи между образами.

• Ввод в систему обучающей информации осуществляется путем внесения данных о парах в такие отношения, как тип связи

$$Link \subset Link - Id \times Link - Type \times Specification, \quad (5)$$

где *Link-Id* – уникальный код типа связи, *Link-Type* – вербальное обозначение типа связи, *Specification* – правила применения типа связи и вопросительное местоимение

$$Inter - Pronoun \subset Pronoun - \\ - Id \times Link - Id \times Pronoun, \quad (6)$$

где *Pronoun-Id* – уникальный код местоимения, *Pronoun* – вербальное обозначение местоимения.

• Вопросительное местоимение между образами пары задается двумя способами. Вначале можно выбрать тип связи из кортежей отношения *Link* (определение, сказуемое, подлежащее, обстоятельство места, обстоятельство времени, дополнение), тогда выбранный тип служит фильтром, и количество возможных местоимений *Inter-Pronoun* уменьшается. С другой стороны, если выбран сначала вопрос как кортеж из *Inter-Pronoun*, то для контроля пользователю автоматически демонстрируется соответствующий ему тип связи из *Link*.

• Пользователю предоставляется возможность выбрать каждое слово пары в меню, составленном из слов текущего предложения *Event*, а затем указать роль соответствующего образа в синтагме (качество объекта, объект, понятие, метод, качество метода). Слова представлены в виде отношения

$$Words \subset Word - Id \times Word \times Bi - I, \quad (7)$$

где *Word-Id* – уникальный код слова, *Word* – собственно слово, *Role-Id* – уникальный код роли слова в синтагме, а роли – в виде отношения

$$Role \subset Role - Id \times Role - Type, \quad (8)$$

где *Role-Type* – вербальное обозначение роли слова.

• С целью привязки выбранного слова пары к образу составляется ранжированный список наиболее схожих внешне слов из уже существующих в словаре образов *Image*. При составлении списка [3] учитываются правила синтаксиса, например, если идет речь о сказуемом, то в список необходимо помещать атрибуты *Method* или *Notion*. Пользователю предоставляется возможность выбрать в меню опцию, при необходимости откорректировать соответствующую статью словаря образов или ввести совершенно новую.

• Завершается формирование базы данных тезауруса образов с помощью отношения *Construct*, в кортежах которого фиксируются особенности использования одинаковых ассоциативных пар образов в различных предложениях:

$$Construct \subset Construct - Id \times Pronoun - \\ Id \times Twice - Id \times Word - Id_1 \times Word - \\ - Id_2 \times Event - Id \times Role - Id, \quad (9)$$

где *Construct-Id* – уникальный код конструкта синтагмы.

При создании новой записи в таблице *Construct* параллельно в соответствующую запись таблицы *Assoc-Twice* обязательно добавляется единица или в поле *Force*, или в поле *Force* ¯ в зависимости от того, в прямом или в обратном порядке проявилась данная с связь.

Реализация операций ассоциативного поиска

Следует отметить, что ценность в предложенной реляционной модели представляют отношения *Image* и *Assoc-Twice*, поскольку они совместно образуют ассоциативную сеть образов и несут главную смысловую нагрузку. Все остальные отношения фактически моделируют ленту событий или долговременную память [2]. Поэтому с целью обеспечения многопользовательского доступа к тезаурусу образов предлагается содержание в одном файле двух таблиц *Image* и *Assoc-Twice* как ядра системы, отдельное его администрирование и организация к нему параллельного доступа *on-Line*.

Действующий прототип тезауруса образов программно реализован на основе технологии *Python + SQLite*, соединяющий возможности языка запросов *SQL* к реляционной базе данных с парадигмами объектно-ориентированного и функционального программирования. На рис. 1 представлена схема данных тезауруса образов, используемая в *SQLite* в соответствии с реляционной моделью (1)÷(9).

Тестовый пример предложенного подхода реализован на основе внесения в систему упрощенного варианта (без предлогов) известного текста А.С. Пушкина «Сказка о рыбаке и золотой рыбке». На рис. 2 представлен внешний вид оболочки обучающей системы, заполненный тестовыми синтагмами. При необходимости удаления синтагмы используется соответствующая ей кнопка «*x*». Для ввода в систему новой синтагмы очередное предложение записывается в верхнем поле и нажимается кнопка «+». После этого система переходит в режим редактирования синтагмы, пример которого показан на рис. 3 для первого предложения текста. В таком режиме каждому слову синтагмы соответствует единственный образ из существующих в словаре *Image* или через ссылку/режим «*Images*» вводится новый образ. Кроме этого, с помощью трех списков в нижней части окна (рис. 3) и набора кнопок «*x*» и «+» реализована функция удаления и создания новых ассоциативных пар синтагмы, причем из левого списка выбирается главное слово пары, из среднего – вопросительное местоимение, а из третьего – подчиненное слово пары.

На рис. 4 представлен пример режима графической интерпретации связей для той же самой первой синтагмы. Переход в данный режим осуществляется с помощью графической ссылки в верхней части окна режима редактирования (рис. 3). Режим графической интерпретации предназначен для визуального контроля правильности задания связей между всеми словами и соответствующими образами синтагмы. Для программной поддержки режи-

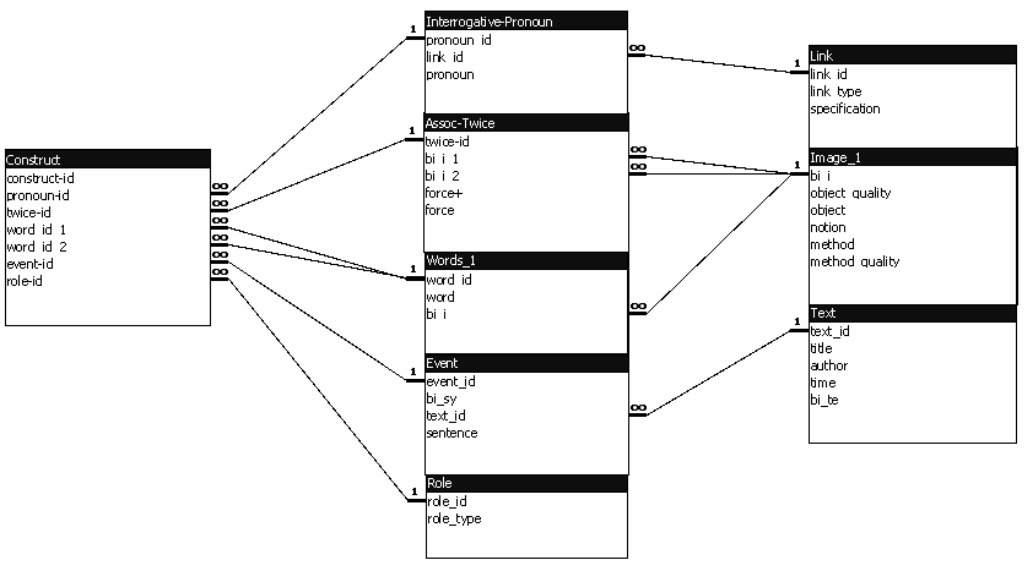

Рис. 1. Схема данных тезауруса образов

ма использованы возможности графической библиотеки *Graphviz*.

Рис. 2. Внешний вид оболочки обучающей системы

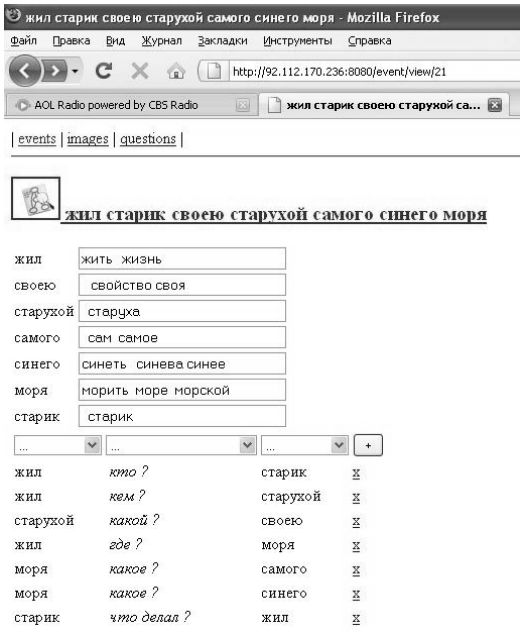

Рис. 3. Режим редактирования синтагмы в оболочке обучающей системы

Рассмотрим реализацию типов ассоциативного поиска [4] в условиях тезауруса образов. Первым из них является известный поиск типа «ИЛИ» – предложенный подход обеспечивает поиск всех словоформ без дополнительного морфологического анализатора.

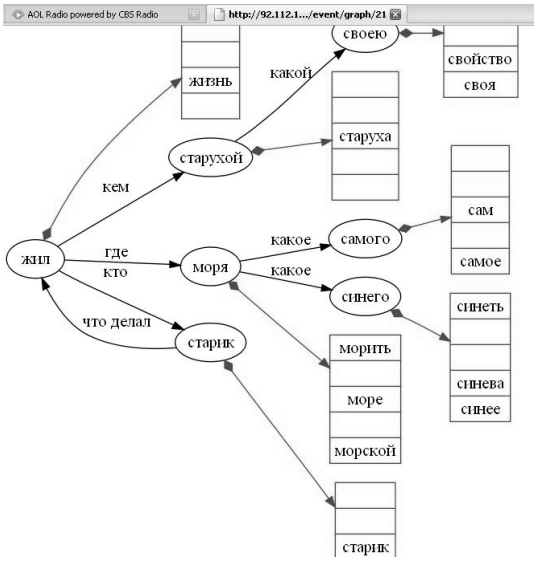

Рис. 4. Режим графической интерпретации связей синтагмы

Реализуется поиск типа «ИЛИ» на основе следующего *SQL*-запроса (код 12 в тестовом примере соответствует образу «старик»):
select syntagma from event where event_id in (*select distinct event_id from construct where word_id_*1 *and word_id_*2 *in* (*select word_id from words where bi_i* = 12)).

На рис. 5 показан результат поиска «ИЛИ» всех синтагм тестового примера, где присутствует образ «старик».

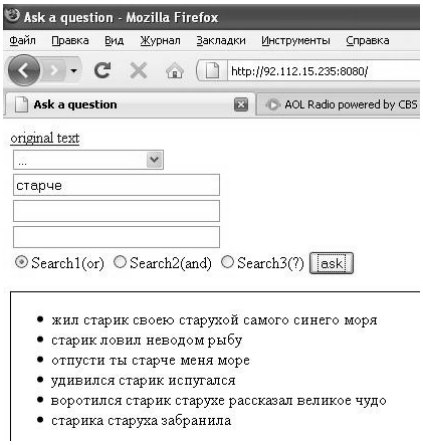

Рис. 5. Поиск всех предложений с образом «старик» в любой словоформе

Предложенная оболочка системы автоматически вставляет в текст *SQL*-запроса код образа, выбранного пользователем в одном из

полей окна поиска (рис. 5). При необходимости поиска типа «ИЛИ» нескольких образов текст запроса незначительно меняется (код 13 в тестовом примере соответствует образу «старуха»):

select syntagma from event where event_id in (*select distinct event_id from construct where word_id_*1 *and word_id_*2 *in* (*select word_id from words where bi_i* = 12 *or bi_i* = 13)).

На рис. 6 показан тестовый результат поиска типа «ИЛИ» для двух образов («старик» и «старуха»).

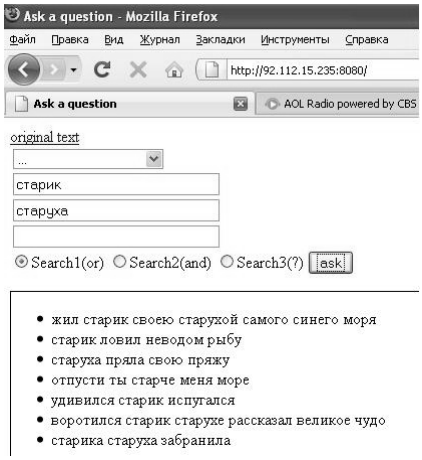

Рис. 6. Поиск всех предложений с образами «старик» или «старуха»

Следующий тип поиска «И» также сохраняет независимость своих результатов от словоформ двух–трех входящих в него образов и запускается переключателем *Search*2(*and*).

Несмотря на формальную близость к первому типу поиска («ИЛИ») поиск «И» гораздо ближе к человеческому мышлению. Так, показанный на рис. 7 тестовый результат поиска типа «И» для тех же двух образов «старик» и «старуха» значительно меньше предыдущего.

Данный тип поиска моделирует ситуацию, когда человек, думая о чем-то своем, рассеянно слушает собеседника. Или слышит текст на малознакомом языке и улавливает смысл не всех слов. В таких случаях на основе нескольких воспринятых слов в его голове непроизвольно возникает воспоминание о событии, где эти образы присутствуют совместно. Применение предложенного подхода обеспечивает появление на выходе системы цитаты из текста в виде той синтагмы, в которой присутствуют входные образы. Формальный алгоритм поиска «И» состоит из следующих шагов:

Ш а г 1. Определить коды образов для каждого входного слова и составить из них множество *Images-List-X*.

Ш а г 2. Аналогично определять множество *Images-List-i* для каждой *i*-й синтагмы из *n* существующих.

Ш а г 3. Отбирать в выходной список *Syntagma-List* те *i*-е синтагмы, для которых

$$Images - List - X \subseteq Images - List - i\,. \quad (10)$$

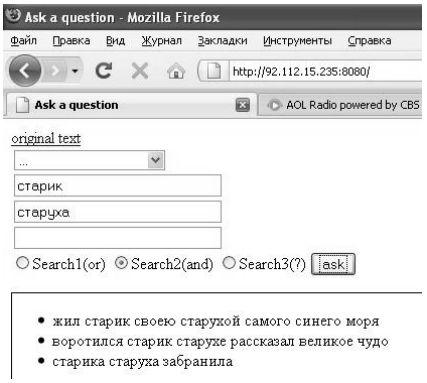

Рис. 7. Поиск всех предложений, в которых есть образы «старик» и «старуха»

Третьим из рассмотренных типов поиска *Search*3(?) будем считать ответ на формальный вопрос, начинающийся вопросительным местоимением *Pronoun*. В окне формы, изображенной на рис. 5÷7, местоимение *Pronoun* можно выбрать из верхнего списка, что исключает использование неизвестных системе вопросов. Результатами данного типа поиска принципиально могут быть: а) пустой список, б) слово, в) часть предложения, г) множество синтагм *Syntagma-List*. Используя особенности предложенного подхода, формальный алгоритм ответа на вопрос представим в виде следующих шагов:

Ш а г 1. Если результат поиска типа «И» *Syntagma-List* пуст, то выходом является вариант а).

Ш а г 2. Иначе организовать цикл по элементам множества *Syntagma-List*:

- разбить очередную синтагму на список составляющих ее пар *Pair-List*;

• проверить в цикле каждую пару из *PairList* с целью нахождения такой из них, в которой главный образ совпадает с образом первого слова, а вопросительное местоимение пары совпадает с вопросительным местоимением *Pronoun*; если соблюдается только первое условие и тип связи *link_id* у обоих вопросительных местоимений совпадает, пару считать найденной;

◊ если поиск пары успешен, то выполняется следующая проверка: является ли подчиненный образ пары главным для каких-либо других пар данной синтагмы:

– если нет, то подчиненный образ пары заложить в вариант б);

– если да, то ответ по варианту в) формировать с помощью рекурсивного алгоритма, «вытягивающего» из синтагмы поддерево образов, подчиненных подчиненному образу пары.

Ш а г 3. Если образ первого слова ни разу не был главным ни в одной из пар, то выходом является вариант г).

На рис. 8 показан пример тестового результата поиска типа *Search*3(?), реализующего вышеизложенный алгоритм с выходом по варианту в).

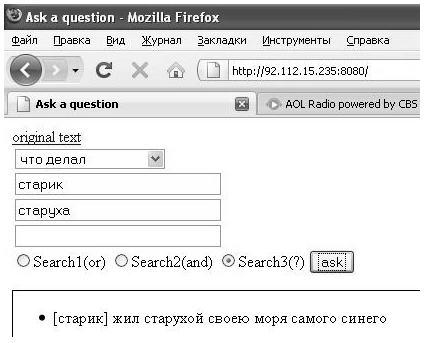

Рис. 8. Поиск типа *Search*3(?) в виде нахождения ответа на вопрос

Заключение. Предложены формальные методы ассоциативного поиска на основе использования реляционной модели тезауруса образов, наполняемого в результате обработки текстов с обучающим материалом. Показаны преимущества и практические результаты последовательного применения в рамках предложенного подхода к моделированию образного мышления трех типов поиска: «ИЛИ», «И» и ответ на вопрос с вопросительным местоимением. Для программной реализации обучающей системы выбрана технология *Python* + *SQLite* + *Graphviz*. Новые возможности предложенного подхода достигаются большей трудоемкостью внесения текстов учебного содержания в систему, однако сравнительно небольшие объемы обучающего контента [5] не позволяют считать такое ограничение критическим.

Перспективным исследованием представляется построение универсальных алгоритмов обработки образных конструкций и дальнейшее развитие на этой основе формальных методов моделирования инсайта, механизма обмена образами в оперативной памяти и других феноменов образного мышления.

1. *Манако А.Ф.*, *Манако В.В.* Електронне навчання і навчальні об’єкти. – К.: ПП «Кажан плюс», 2003. – 334 с.
2. *Бісікало О.В.* Алгебраїчна модель лінгвістичного процесора // Информационные технологии в управлении сложными системами: Сб. докладов и тезисов Междунар. науч.-практ. конф. (Днепропетровск, 22–23 мая 2008 г.). – Днепропетровск: ИТМ НАНУ и НКАУ, 2008. – С. 23–24.
3. *Bisikalo O.* Knowledge base of teaching system construction supported by creative thinking model // Third Intern. Conf. «New Information Technologies in Education for All: e-education», Proc. (1–3 Oct. 2008). – Kiev, Akademperiodika, 2008. – P. 413–421.
4. *Бісікало О.В.* Класифікація образного пошуку // Тези доп. Першої міжнар. наук.-тех. конф. «Інтелектуальні системи в промисловості і освіті – 2007», 7–9 лист. 2007 р. – Суми, 2007. – С. 14–15.
5. *Бісікало О.В.* Проектування електронного підручника на основі формалізації пізнавальної діяльності людини // Зб. наук. пр. «Перспективні технології навчання та освітні простори». – Київ, МННЦ ІТтаС, 2007. – **1**. – С. 179–190.

© О.В. Бисикало, 2009