Для реализации анализируемых алгоритмов были разработаны программы с использованием архитектуры клиент-серверного приложения.

Целью написания программ являлось сравнение скорости обучения в локальной сети в зависимости от длины ключа и в дальнейшем – устойчивости каждого алгоритма к геометрическим атакам.

Программы написаны в среде Microsoft Visual Studio 2005 как Win32 console application. В окне клиентской части программы, реализующей методы TPCM и TPM, задается количество шагов и адрес сервера, к которому производится подключение клиента. Интерфейс программы, реализующей метод BPM, аналогичен интерфейсу, приведенному выше. Отличие заключается в том, что в BPM не надо задавать количество шагов обучения, так как особенностью метода является то, что он сам «решает», когда нейронные сети полностью синхронизировались.

При проведении тестирования (на время обучения) получены следующие результаты (таблица).

Статистика результатов проведенных испытаний

Алгоритм обучения	Время синхронизации (с) двух сетей при получении ключа длиной		
	48 символов	66 символов	132 символа
TPM	22,176	28,393	40,3
TPCM	1 800,27	2 270,5	3 287,88
BPM	35 770,2	49 084,1	98 368,23

Заключение. Как видно из проведенного анализа и данных в таблице, преимуществом алгоритмов обучения сетей TPM и TPCM является скорость обучения. Однако их главный недостаток заключается в том, что обучающиеся нейронные сети наперед «не знают» количества шагов, за которые они обучатся, и пользователи практически вынуждены задавать их вручную. Но с увеличением длины ключа количество шагов растет, и, следовательно, пользователи должны каждый раз подбирать это количество. Также время обучения по данным методикам зависит от пороговых значений L. В данном примере принято, что L изменяется в диапазоне от –3 до 3, и, соответственно, весовые коэффициенты принимают всего 7 различных значений. Преимуществом алгоритма BPM является то, что сети «знают», когда они полностью обучились. Кроме того, данный алгоритм обучения использует алгебру дробных чисел, и точность после запятой в данном примере достигает 4 знака (это числа в диапазоне от 0,0001 до 0,9999 с шагом 0,0001), т. е. 10 000 различных значений. Точность алгоритма BPM зависит от функции, которая применяется для расчета ошибки: чем «грубее» функция, тем больше шаг и, следовательно, меньше точноть. Но достоинством такого подхода является меньшее время, затрачиваемое на обучение нейронных сетей. При применении более точной функции время сходимости увеличивается, однако точность получаемого результата также растет.

Литература

1. Плонковски, М. Криптографическое преобразование информации на основе нейросетевых технологий / М. Плонковски, П. П. Урбанович // Труды БГТУ. Сер. VI, Физ.-мат. науки и информатика. – 2005. – Вып. XIII. – С. 161–164.

2. Галушкин, А. И. Синтез многослойных систем распознавания образов / А. И. Галушкин. – М.: Энергия, 1974. – С. 25–27.

3. Werbos, P. J. Beyond regression: New tools for prediction and analysis in the behavioral sciences / P. J. Werbos // Ph.D. thesis. – Harvard University, Cambridge, MA, 1974. – P. 87–89.

4. Rumelhart, D. E. Learning Internal Representations by Error Propagation. In: Parallel Distributed Processing / D. E. Rumelhart, G. E. Hinton, R. J. Williams. – 1986. – Vol. 1. – P. 318–362.

5. Барцев, С. И. Адаптивные сети обработки информации / С. И. Барцев, В. А. Охонин. – Красноярск: Ин-т физики СО АН СССР, 1986. – С. 102–109.

Поступила в редакцию 31.03.2010

СИСТЕМНЫЙ АНАЛИЗ И ОБУЧАЮЩИЕ СИСТЕМЫ

УДК 512.8

Н. И. Гурин, доцент (БГТУ); О. В. Герман, доцент (БГТУ)

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗАТОР ЗАПРОСОВ К БАЗЕ ЗНАНИЙ МУЛЬТИМЕДИЙНОГО ЭЛЕКТРОННОГО УЧЕБНИКА

Создана активная обучающая среда для электронного учебника, которая реализует эффект присутствия виртуального преподавателя с непрерывным контролем приобретаемых знаний. Активная обучающая среда использует семантическую сеть учебных объектов и проводит непрерывный контроль знаний. Семантический анализатор позволяет получить ответ на любой вопрос по изучаемой дисциплине, сформулированный в рамках определяемых для него правил. Семантический анализатор работает независимо от наполнения базы знаний и может быть использован для организации активной обучающей среды электронного учебника по любой дисциплине.

It is developed the intellectual analyzer of questions to knowledge base of the electronic textbook with active training environment witch realizes effect of presence of the virtual teacher. The active training environment execute semantic network of educational objects and carry out the continuous control of knowledge. The semantic analyzer allows to receive the answer to any question on the studied discipline, formulated within the rules defined for it. This analyzer works irrespective of filling of the knowledge base and can be used for active training environment of the electronic textbook on any discipline.

Введение. Электронный учебник с активной обучающей средой [1, 2] реализует эффект присутствия виртуального преподавателя, который предоставляет необходимую помощь, если обучаемый затрудняется или не может самостоятельно разобраться с учебным материалом. Для функционирования такой активной обучающей среды, прежде всего, строится семантическая сеть учебных объектов электронного учебника, которая является опорным компонентом для функционирования всей системы. Затем задаются критерии, по которым определяются реакции обучающей среды на действия обучаемого. Заключительным этапом работы по созданию активной обучающей среды является создание базы знаний для электронного учебника и разработка семантического анализатора запросов к ней. При этом семантический анализатор базы знаний является ключевым компонентом активной обучающей среды, используя который студент может получить ответ на любой свой вопрос по изучаемой дисциплине, конечно, заданный в рамках правил, принятых для семантического анализатора. С учетом последнего фактора активная обучающая среда электронного учебника фактически становится экспертной системой, позволяющей как оценить действия обучаемого, так и оказать ему необходимую помощь в ходе изучения дисциплины.

Учебный материал мультимедийного электронного учебника является в основном текстовым документом, в котором наряду с текстом содержатся также графические объекты, формулы, таблицы, диаграммы, анимации, аудио- и видеозаписи и т. п., обращение к которым в учебнике сводится к обработке имен соответствующих файлов, также являющихся текстовой информацией. С текстом можно выполнять различные задачи, в том числе интеллектуальные. Под интеллектуальной задачей понимается задача, которая характеризуется следующим перечнем свойств:

1) задача не имеет хорошего алгоритма решения (с математической точки зрения);

2) задача плохо формализована;

3) имеются эксперты по задаче;

4) задача характеризуется неполнотой или ненадежностью сведений;

5) задача связана с огромным множеством вариантов и альтернатив.

Даже часть из этого перечня позволяет рассматривать задачу как интеллектуальную. В данной работе исследуются именно интеллектуальные задачи обработки текстов.

Основная часть. При работе с текстом как с базой знаний мы имеем дело с текстом как семантическим объектом. Это значит, что текст выступает в качестве базы знаний. Для такой базы знаний необходимо организовать соответствующий интерфейс, предполагающий обработку семантических и поисковых вопросов, организацию контекстной помощи и тестирование

знаний учащихся. Семантический вопрос требует выдачи точечного ответа, в отличие от поискового вопроса, ответом на который является список ссылок на различные релевантные (подходящие по смыслу) места в тексте.

Чтобы обеспечить семантическую обработку вопросов, текст необходимо представить в формализованном виде. В принципе имеется несколько способов формального представления текста для семантической обработки. Можно указать два основных. Во-первых, это *семантическая сеть*. Во-вторых, это *семантическая база данных с интерфейсом*, содержащим машину логического вывода.

Как правило, любой текст можно охарактеризовать набором ключевых слов, которые этот текст представляют и играют роль смысловых слов текста. Поиск ключевых слов – это тоже важная техническая задача, на которой следует остановиться. Когда имеется множество текстовых документов со своими ключевыми словами, то возникает проблема построения поискового дерева. Такое дерево позволяет отыскивать требуемый текстовый документ с незначительными временными издержками. Можно выполнять поиск и без дерева на основе известных статистических методов, использующих, например, теорему Байеса.

Для определения множества ключевых слов первоначально рассматривают только существительные (прилагательные, глаголы, местоимения, наречия не рассматривают). Подсчитываются частоты или частости слов (сколько раз каждое слово вошло в текст). Поскольку слова входят в текст с разными окончаниями, то поступают следующим образом: выбрасывают из слова гласные и считают два слова совпадающими, если они достаточно похожи друг на друга. При этом короткие слова должны быть похожими друг на друга в большей степени, чем длинные. Например, рассмотрим два слова: *«текстом»* и *«текста»*. После отбрасывания гласных получим соответственно: *«текстм»* и *«текст»*. Оценим степень θ совпадения этих слов по следующей формуле: θ = (размер совпадающей части) / (размер слова-образца).

В нашем примере размер совпадающей части составляет 5 (число букв в слове *«текст»*). Максимальный размер слова – 6 (*«текстм»*). Таким образом, степень совпадения составляет 5/6 ≈ 0,83. Выделение совпадающей части в целом не вызывает проблем, если слова написаны без ошибок. Обработка слов с потенциальными ошибками требует специфических методов. Наконец, остается еще вопрос учета длины слов. Чем длиннее слова, тем меньше требуется допустимый процент совпадения. Этот процент можно установить исходя из следующей эмпирически составленной таблицы.

Таблица совпадений

Размер слова	Допустимое число неверных символов
До 5	0
5–6	1
7–8	2
9–10	3
Больше 10	4

Таким образом, степень совпадения колеблется в интервале 0,66–0,85, что на практике является вполне приемлемым.

Итак, пусть определены частоты встречаемости слов. Размещаем слова по убыванию частот и оставляем примерно $\sqrt{N}$ слов с наибольшими значениями частот (здесь N – число всех слов). Эти слова несут основную семантическую нагрузку текста. Они играют роль паспорта текстового документа.

При отыскании документа задают множество ключевых слов $K = \{k_1, k_2, ..., k_n\}$. Задача состоит в том, чтобы из множества текстовых документов выбрать те, паспорта которых имеют достаточно много совпадений со словами из $K = \{k_1, k_2, ..., k_n\}$. Пусть имеется множество кластеров документов, причем каждый документ описывается набором ключевых слов. Зададим поисковый набор ключевых слов и выясним, к какому кластеру относится данный поисковый набор. Для ясности будем использовать следующие два кластера:

$$K_1 = <D_1 = \{k_1, k_3, k_4, k_5\}, D_2 = \{k_1, k_3, k_6\},$$
$$D_3 = \{k_2, k_3, k_4, k_5\}, D_4 = \{k_3, k_4, k_8\}>$$

и

$$K_2 = <D_1 = \{k_1, k_8, k_9, k_{10}\}, D_2 = \{k_7, k_9, k_{11}\},$$
$$D_3 = \{k_2, k_3, k_7, k_{10}\}, D_4 = \{k_1, k_9, k_{12}\}>.$$

Пусть поисковый набор $Z = k_1, k_3, k_5, k_8$. Возьмем документ D_1 первого кластера и найдем, сколько слов совпадает у него с поисковым набором 3. Процент совпадения составит 100% · 3/4 = 75%. Аналогично процент совпадения по другим документам из первого кластера будет: 66, 50, 66%. Вычислим теперь средневзвешенный процент совпадения по всему кластеру (≈57%).

Для второго кластера получим процент совпадения 21%. Очевидно, поисковый образец следует отнести к первому кластеру.

Алгоритм обработки запросов к текстовой базе знаний реализуется следующим образом:

– идентифицируется тема, к которой относится запрос;

– идентифицируется блок текста, содержащий ответ на вопрос;

– из блока текста извлекается требуемая часть для ответа на вопрос.

Принципиально оба указанных шага могут быть выполнены с позиций оценки расстояния между двумя текстовыми документами, либо расстояния между текстовым документом и вопросом. Приведем иллюстрацию на примере электронного учебника с активной обучающей средой, разработанного по учебному пособию [3]. Пусть вопрос имеет такой вид: *«что такое химический потенциал?»*.

Пусть выбран текстовый документ, состоящий из предложений следующего вида:

1) «химический потенциал – это величина заряда ядра»;

2) «химический потенциал имеет размерность Дж/моль»;

3) «химический потенциал зависит от давления»;

4) «химический потенциал не зависит от массы»;

5) «химический потенциал зависит от фазы вещества»;

6) «химический потенциал зависит от температуры»;

7) «химический потенциал определяет направление самопроизвольного переноса вещества»;

8) «перенос вещества выравнивает химический потенциал»;

9) «химический потенциал имеет одинаковую величину, когда вещество находится в равновесии»;

10) «химические потенциалы равны, когда система находится в термодинамическом равновесии»;

11) «химический потенциал равен работе для перевода ядра в неактивное состояние»;

12) «фаза – это однородное по составу вещество»;

13) «фазой называется однородное по составу вещество»;

14) «фаза определяет устойчивое состояние вещества».

Из этого файла отбираются только предложения 1–7, 9–11, как имеющие наибольшее сходство с текстом вопроса. Для извлечения окончательного ответа следует выполнить грамматический разбор каждого из выбранных предложений 1–7, 9–11. Реализация грамматического разбора преследует цель выявить:

– вопросное слово;

– структурные составляющие предложения;

– структурные составляющие вопроса и их соответствие;

– степень соответствия структурных составляющих предложения и вопроса с учетом вопросного слова.

В качестве иллюстрации на рис. 1 приведен пример обработки вопроса: *«чем характеризуется ЭДС?»*.

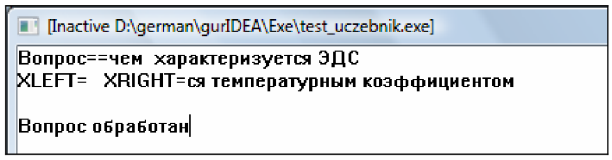

Рис. 1. Окно вывода ответа
на вопрос: «чем характеризуется ЭДС?»

Рассмотрим другое предложение: *«ЭДС зависит от температуры по линейному закону $E = a + bT$»*.

Структура этого предложения такова:

1) <Блок существительного> = *«ЭДС»*;

2) <Блок глагола> = *«зависит»*;

3) <Блок существительного> = *«от температуры по линейному закону $E = a + bT$»*.

К этому предложению можно составить следующие вопросы:

– что зависит от температуры;

– от чего зависит ЭДС;

– как зависит ЭДС от температуры;

– по какому закону зависит ЭДС.

Очевидно, что обработка вопроса прямым образом связана с вопросным словом и структурной организацией предложения. Мы должны выделять субъект, объект и действие. Субъект – это то лицо или предмет, которое производит действие. Объект – это то лицо или предмет, на которое направлено действие. Вопросы к субъектному блоку начинаются со слов:

1) *кто* (производит действие);

2) *что* (производит действие);

3) *какой* (-ая, -ое) производит действие;

4) *что такое…*;

5) *про что* (говорится)…;

6) *о чем* (говорится)…

Например, на вопрос: *«что такое ЭДС?»* получим ответ, представленный на рис. 2.

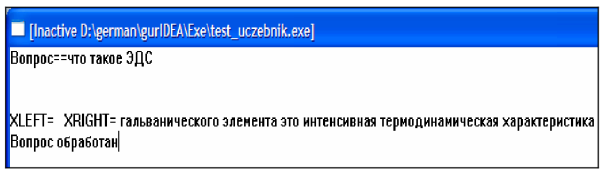

Рис. 2. Окно вывода ответа
на вопрос: «что такое ЭДС?»

Заключение. Реализация смысловых запросов к текстовой базе знаний позволяет

существенно повысить интеллектуальный уровень электронных учебников, интерактивных поисковых систем и систем обучения с виртуальным преподавателем. Полученные практические результаты показывают, что системы подобного класса могут быть реализованы достаточно эффективно, а уровень интеллектуальности учебника связан не столько с механизмом вывода, сколько с мощностью текстовой базы.

Литература

1. Гурин, Н. И. Активный контроль знаний в электронной обучающей системе / Н. И. Гурин, Т. В. Мицкевич // Технологии электронного обучения в современном ВУЗе: материалы Междунар. науч.-техн. конф., Минск, 13–15 мая 2008 г. / Гос. ин-т управления и соц. технологий БГУ. – Минск, 2008. – С. 129–130.

2. Гурин, Н. И. Организация структуры электронной обучающей системы с активным контролем приобретаемых знаний / Н. И. Гурин, О. В. Герман // Труды БГТУ. Сер. VI, Физ.-мат. науки и информатика. – 2009. – Вып. XVII. – С. 107–110.

3. Дудчик, Г. П. Равновесная электрохимия. Электроды и гальванические элементы / Г. П. Дудчик, И. М. Жарский. – Минск: БГТУ, 2000. – 160 с.

Поступила в редакцию 31.03.2010