

*Е. Г. Малькова, Ю. Е. Усачёв, Е. Н. Яшина***АВТОМАТИЗАЦИЯ ПРОЦЕССА РЕЦЕНЗИРОВАНИЯ
УЧЕБНО-МЕТОДИЧЕСКИХ ИЗДАНИЙ
НА ОСНОВЕ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТА***E. G. Malkova, Yu. E. Usachyov, E. N. Yashina***AUTOMATIZATION OF THE REVIEWING PROCESS
OF TEACHING MATERIALS
ON THE BASIS OF SEMANTIC TEXT ANALYSIS**

Рассматривается автоматизация процесса рецензирования учебно-методических изданий. Используется подход, основанный на создании в автоматическом режиме иерархической структурной модели текста документа с помощью комплексного смыслового анализатора текста – программы Text Analyst. Приводится методика определения степени соответствия учебного пособия федеральным государственным стандартам высшего профессионального образования. Описана реализация системы рецензирования и приводятся результаты ее апробации.

Ключевые слова: модель текста, семантическая сеть, рецензирование, дидактическая единица.

The automatization of the reviewing process of teaching materials is considered in the paper. The approach based on creation of a hierarchical structural model of the document text in an automatic mode by means of the complex semantic analyzer of the text –Text Analyst programs, is used. The technique of determination of the correspondence level of the manual to the federal state educational standard of the higher vocational training is shown. The implementation of the reviewing system is described, and the results of its approbation are given.

Key words: text model, semantic network, reviewing, didactic unit.

Введение

Одним из основных компонентов методического обеспечения учебного процесса является учебно-методическое издание по дисциплине. Его контент и качество – значимый фактор, влияющий на уровень подготовки специалистов. Содержание учебно-методических изданий в первую очередь должно соответствовать федеральному государственному образовательному стандарту высшего профессионального образования (ФГОС ВПО). Актуальной является проблема создания автоматизированной системы, способной выполнять часть функций экспертов при анализе смыслового содержания учебного пособия путем определения наличия в нем дидактических единиц, содержащихся в рабочей программе.

В статье описывается автоматизация процесса рецензирования учебно-методических изданий, позволяющего качественно и эффективно проводить анализ смыслового содержания учебных пособий на предмет соответствия его требованиям ФГОС ВПО. Это позволит сократить трудоёмкость выполнения и повысить качество экспертной оценки учебных пособий.

Постановка задачи

Имеется ФГОС ВПО специальности, содержащий описание дисциплины, в котором приведены основные дидактические единицы. По этому описанию разработана рабочая программа дисциплины.

Необходимо разработать систему автоматизированного рецензирования учебных пособий, которая должна выделять основные понятия, содержащиеся в тексте учебного пособия. Для выделенных понятий система определяет семантические отношения между ними и оформляет модель взаимосвязи понятий в форме семантической сети. Система должна выполнять анализ семантической сети понятий с целью определения наличия в учебном пособии дидактических единиц, указанных в рабочей программе дисциплины.

Подход к решению

Рассматриваемый нами подход основан на автоматическом извлечении знаний из текста учебного пособия с помощью смыслового анализатора текста. Система генерирует семантический (смысловой) «портрет» документа в виде семантической сети, содержащей основные понятия и описывающей отношения между ними. Семантическая сеть отображает суть учебного по-

собия. Она включает только значимые для данного текста понятия. Для определения наличия некоторой дидактической единицы в тексте учебного пособия необходимо найти соответствующие ей понятия в сгенерированной семантической сети с сохранением их следования в дидактической единице.

В настоящее время для реализации данного подхода существует достаточно много инструментов текстомайнинга, таких как WordStat, Aerotext, Businessobjects Text Analysis, Attensity Text Analytics suite и др. [1].

Сравнение этих программ показало, что в качестве инструментального средства анализа текста учебного пособия следует выбрать комплексный смысловой анализатор текста Text Analyst [2], т. к. он позволяет решить поставленные задачи, настраиваясь на конкретную предметную область с помощью пользовательских словарей, и является наиболее доступным. Кроме того, задавая различные веса понятиям и смысловым связям, можно анализировать документы с различным уровнем детализации.

Формальная модель текста документа

В простейшем случае структуру S совокупности знаний можно определить следующим образом:

$$S = \{M, F\},$$

где M – множество всех понятий данной совокупности знаний; F – отношение «смысловая связь».

Бинарное отношение F на множестве всех понятий данной совокупности знаний M является антирефлексивным, симметричным и антитранзитивным.

Определение 1. Понятие $a \in A$ находится в отношении «смысловая связь» с каждым из понятий множества $B_a \subset A$, которое сочетается по смыслу в тексте (часто используются в сочетании) с понятием a .

Каждому понятию множества $a \in A$ во взаимно однозначное соответствие поставлено множество B_a , удовлетворяющее следующим условиям:

$$\forall a \in A \exists B_a \mid B_a \subset M \wedge (\forall b \in B_a, aFb).$$

Базовое множество понятий M определяется по формуле

$$M = A \cup \bigcup_{i=1}^m B_i, \quad m = |A|.$$

В качестве формальной модели структуры знаний можно использовать семантическую сеть, определяемую в виде ориентированного графа

$$G = (E, V), \quad (1)$$

где E – множество вершин, поставленное во взаимно однозначное соответствие с множеством понятий; V – множество ориентированных дуг; $v = (x, y) \in V$; $x \in E \wedge y \in E$ – дуга выходит из вершины, соответствующей основному понятию a , и входит в вершину, соответствующую понятию, которое сочетается по смыслу в тексте (часто используются в сочетании) с понятием a .

Формальная постановка задачи

Имеется основной документ, рабочая программа, составленная согласно ФГОС специальности. Существует множество $A = \{a_1, a_2, \dots, a_n\}$ основных понятий этого документа. В рабочей программе существует множество дидактических единиц $D = \{d_1, d_2, \dots, d_n\}$, каждая из которых включает несколько основных понятий документа.

Определение 2. Дидактическая единица $d_i \in D$ находится в отношении «Дидактическая единица» с каждым из понятий множества $B_i \subset M$, которые входят в ее состав. На множестве B_i задано отношение следования.

Каждой дидактической единице $d_i \in D$ можно поставить в соответствие модель в виде графа

$$G_{d_i} = (E_{d_i}, V_{d_i}), \quad (2)$$

где E_{d_i} – множество вершин, поставленное во взаимно однозначное соответствие с множеством понятий дидактической единицы; V_{d_i} – множество ориентированных дуг; $v = (x, y) \in V_{d_i}$; $x \in E_{d_i} \wedge y \in E_{d_i}$ – дуга выходит из вершины x , соответствующей основному понятию $a \in B_i$, и входит в вершину y , соответствующую понятию $b \in B_i$, которое следует за понятием a в дидактической единице.

Для учебного пособия разработана формальная модель в виде семантической сети, представленной в виде ориентированного графа (1).

Необходимо в графе (1) найти подграфы, изоморфные графам (2) каждой дидактической единицы. Если для всех дидактических единиц найдены такие подграфы, то можно сделать вывод, что документ (учебное пособие) соответствует федеральному государственному образовательному стандарту высшего профессионального образования. Если найдутся не все подграфы, то необходимо оценить степень соответствия учебного пособия ФГОС ВПО. Очевидно, это будет соотношение найденных дидактических единиц к общему количеству дидактических единиц.

Метод оценки степени соответствия

При нахождении подграфа, изоморфного данной дидактической единице, возможно частичное совпадение. Степень совпадения определяется по формуле

$$Z_{d_i} = T_n / T_o,$$

где Z_{d_i} – степень совпадения дидактической единицы с текстом документа; T_n – количество найденных понятий дидактической единицы d_i ; T_o – общее количество понятий в составе дидактической единицы d_i .

Степень соответствия документа (учебного пособия) S определяется по формуле

$$S = \sum_{i=1}^n Z_{d_i} / |D|. \quad (3)$$

Пример. Имеется фрагмент семантической сети (рис. 1, табл.) и следующие дидактические единицы:

- 1) интеллектуальные информационные системы;
- 2) нейронные сети;
- 3) реализация экспертных систем;
- 4) базы знаний.

Фрагмент семантической сети

Родитель	Подчиненный
a_1 – интеллектуальных	a_2 – информационных
	a_3 – интерфейс
	a_4 – нейронные
	a_5 – реализация
	a_6 – принципы
a_2 – информационных	a_7 – методы
	a_8 – системы
	a_9 – интеллектуальных информационных
	a_2 – информационных
	a_1 – интеллектуальных
a_4 – нейронные	a_{10} – сети
	a_{11} – функции
	a_{12} – человеческого
	a_{13} – мозга
	a_{14} – искусственного
	a_1 – интеллектуальных

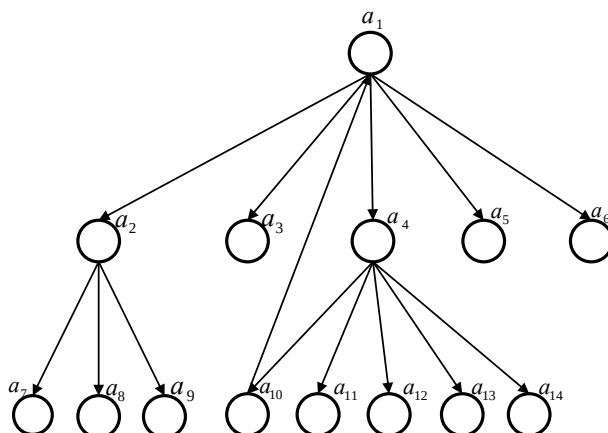


Рис. 1. Фрагмент семантической сети

В результате поиска во фрагменте семантической сети были выделены подграфы, представленные на рис. 2–4.

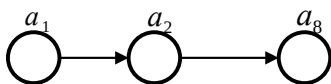


Рис. 2. Дидактическая единица «Интеллектуальные информационные системы»

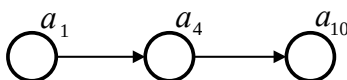


Рис. 3. Дидактическая единица «Нейронные сети»

Был обнаружен также подграф, частично изоморфный дидактической единице «Реализация экспертных систем».

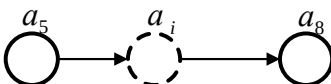


Рис. 4. Дидактическая единица «Реализация экспертных систем»

Подграф, изоморфный дидактической единице «Базы знаний», найден не был. Для каждой дидактической единицы подсчитали степень их совпадения в документе:

$$Z_{d_1} = 3/3 = 1,$$

$$Z_{d_2} = 2/2 = 1,$$

$$Z_{d_3} = 2/3 = 0,67,$$

$$Z_{d_4} = 0/2 = 0.$$

По формуле (3) найдена степень соответствия учебного пособия рабочей программе:

$$S = (1 + 1 + 0,67 + 0) / 4 = 0,6675.$$

Реализация автоматизированной системы рецензирования учебно-методических изданий

Разработанная нами автоматизированная система рецензирования учебно-методических изданий (АСРУМИ) состоит из комплексного смыслового анализатора текстов TextAnalyst, модуля подготовки словаря TANew и модуля рецензирования.

Схема анализа текстов документов с применением АСРУМИ приведена на рис. 5.

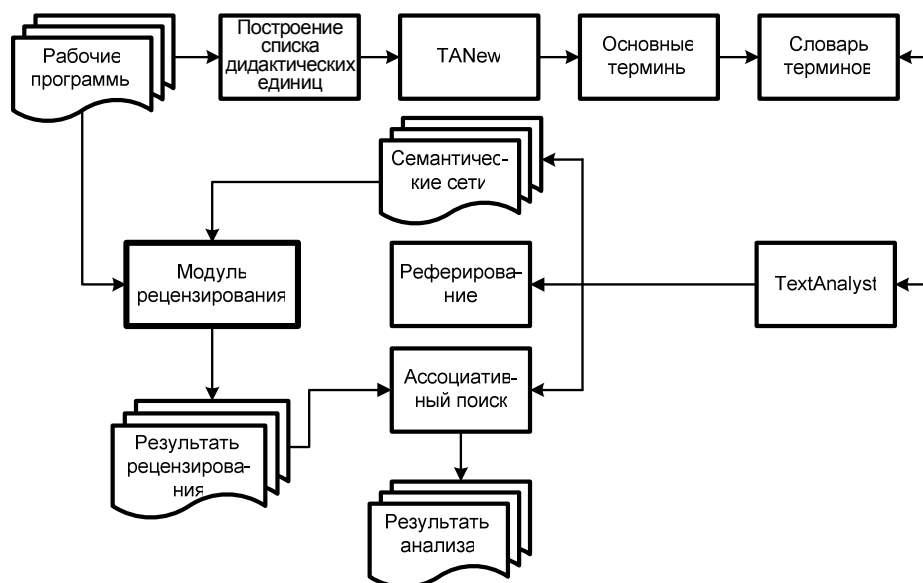


Рис. 5. Схема анализа текстов документов

Для апробации системы было проведено сравнение рабочей программы и учебного пособия по дисциплине «Интеллектуальные информационные системы». В качестве основного документа использовалась рабочая программа дисциплины. При рецензировании была использована методика, описанная в [3].

На первом этапе средствами АСРУМИ из рабочей программы были выделены дидактические единицы и сформирован файл DE.txt, их содержащий.

На втором этапе средствами модуля TANew были подготовлены дополнения к словарю «normal_rus.dic» системы TextAnalyst. Понятия, входящие в состав дидактических единиц, являются словами-предпочтениями, и при анализе отождествляются все слова, имеющие ту же корневую основу, что и основная форма. Из файла, сформированного на предыдущем этапе, были загружены дидактические единицы, после чего автоматически произошло их разбиение на понятия. Далее из полученного списка с помощью соответствующих команд меню были удалены все повторения и произведена сортировка «по имени». Слова-предпочтения были сохранены в файле для последующей работы.

Далее в редакторе словарей VocEdit к основному словарю были присоединены сформированные дополнения и результат объединения был установлен в TextAnalyst для анализа текстов.

На третьем этапе для получения корректных результатов была произведена настройка системы:

- для построения семантической сети понятий и тематической структуры выбраны термины и связи с весом не менее 75 и 1 соответственно;
- для формирования реферата выбраны предложения с весом не менее 90;
- для формирования результатов поиска выбраны термины с весом не менее 1.

Выполнено автоматическое построение семантической сети понятий учебного пособия с использованием комплексного смыслового анализатора текстов TextAnalyst. Результатом построения является семантическая сеть и тематическая структура текста учебного издания. Количество строк семантической сети составило 3 186.

На четвертом этапе средствами АСРУМИ было проведено сопоставление рабочей программы и учебного издания. В результате проведения сопоставления из 77 дидактических единиц, содержащихся в рабочей программе, по семантической сети было построено 58. По формуле (3) была рассчитана степень соответствия учебного пособия образовательным стандартам, которая составила 0,75.

Результаты проведенного сопоставления формируются средствами разработанной прикладной программы (рис. 6).

Рецензирование учебных пособий

Выберите рабочую программу для проведения рецензирования:

☒ Программа дисциплины

D:\DIPLOM\ASPPA_UI\RabProgramma\Sист_интел_дн.rtf

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ФЕДЕРАЛЬНОЕ АГЕНСТВО ПО ОБРАЗОВАНИЮ
ПЕНЗЕНСКАЯ ГОСУДАРСТВЕННАЯ ТЕХНОЛОГИЧЕСКАЯ АКАДЕМИЯ
Факультет «Институт образовательных технологий»

Утверждаю
Декан факультета
«Институт образовательных технологий»

Люсов В.Н.
«___» _____ 2008 г.

ПРОГРАММА ДИСЦИПЛИНЫ

Выберите файл, содержащий семантическую сеть понятий, построенную TextAnalyst:

☒ Сеть понятий TextAnalyst

D:\DIPLOM\ASPPA_UI\ExportBase\ExportBase1.txt

Родитель, Частота, Вес, Подчиненный
аргументов,23,30,=====



Обработать

Отмена

Рецензия

Найденные дидактические единицы (ДЕ):

- Параметризация решаемых задач
- Концептуализация проблемной области
- Структурная модель
- Функциональная модель
- Поведенческая модель
- Реализация экспертной системы
- Развитие прототипа до промышленной экспертной системы
- Оценка системы
- Сопровождение системы
- Особенности экспертных систем экономического анализа
- Рейтинговый метод
- Метод классификации ситуации

Ненайденные дидактические единицы (ДЕ):

- Основные секции программы
- Секция domains
- Секция predicates
- Секция clauses
- Секция goal
- Отсечение
- Рекурсия
- Этапы проектирования экспертной системы: идентификация, концептуализация, формализация
- Участники процесса проектирования: эксперты, инженеры по знаниям, конечные пользователи
- Тестирование
- Внешний и внутренний экономический анализ предприятия
- Метод последовательной декомпозиции

Общее количество ДЕ, выделенных из рабочей программы: 77

Количество ДЕ, найденных в учебном издании: 58

Количество ДЕ, не найденных в учебном издании: 198

Рис. 6. Результат проведения сопоставления

В результате анализа были выявлены дидактические единицы, требующие доработки и более детального представления в учебном издании.

Заключение

В настоящее время задача рецензирования учебных пособий решается исключительно специалистами высокой квалификации, требует больших временных затрат и всегда содержит субъективный фактор. Разработанная система даст возможность в автоматическом режиме осуществлять рецензирование содержания печатных и электронных учебных пособий дисциплин с целью определения их соответствия требованиям ФГОС ВПО, учебных программ и руководящих документов, что позволит повысить качество рецензирования и уменьшить его трудоемкость.

СПИСОК ЛИТЕРАТУРЫ

1. Бельский А. Текстмайнинг. Извлечение информации из неструктурированных текстов // КомпьютерПресс. – 2008. – № 10. – С. 174–179.
2. Харламов А. А. Автоматический структурный анализ текстов // Открытые системы. – 2002. – № 10. – С. 16–22.
3. Автоматизированный анализ соответствия образовательных стандартов и профессиональных требований / Е. В. Жаркова, В. В. Пикулин, Ю. Е. Усачёв, Е. Н. Яшина // Проблемы информатики в образовании, управлении, экономике и технике: сб. ст. X Междунар. науч.-техн. конф. – Пенза: Приволжский Дом знаний, 2010. – С. 125–127.

Статья поступила в редакцию 10.11.2011

ИНФОРМАЦИЯ ОБ АВТОРАХ

Малькова Екатерина Георгиевна – Пензенская государственная технологическая академия; студентка, специальность «Прикладная информатика в экономике»; n.katena89@mail.ru.

Malkova Ekaterina Georgievna – Penza State Technological Academy; Student, Speciality "Applied Informatics (in Economics)"; n.katena89@mail.ru.

Усачёв Юрий Евгеньевич – Пензенская государственная технологическая академия; канд. техн. наук, доцент; кафедра «Прикладная информатика»; ujel@pgta.ru.

Usachyov Yuriy Evgenievich – Penza State Technological Academy; Candidate of Technical Science, Assistant Professor, Department "Applied Informatics"; ujel@pgta.ru.

Яшина Елена Николаевна – Пензенская государственная технологическая академия; студентка, специальность «Прикладная информатика в экономике»; lena_shpilka@mail.ru.

Yashina Elena Nickolaevna – Penza State Technological Academy; Student, Speciality "Applied Informatics (in Economics)"; lena_shpilka@mail.ru.