На правах рукописи

БИДУЛЯ Юлия Владимировна

МЕТОДЫ И АЛГОРИТМЫ СМЫСЛОВОГО ОПИСАНИЯ КОНТЕНТА В СИСТЕМАХ ТЕСТИРОВАНИЯ

10.02.21 – Прикладная и математическая лингвистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата филологических наук

Тюмень - 2011

Работа выполнена на кафедре информационных систем Института математики, естественных наук и информационных технологий ФГБОУ ВПО Тюменский государственный университет.

Научный руководитель доктор технических наук, профессор
ИВАШКО Александр Григорьевич

Официальные оппоненты: доктор технических наук, профессор
ЗАХАРОВ Александр Анатольевич

кандидат филологических наук
БАБИНА Ольга Ивановна

Ведущая организация: **ФГБОУ ВПО Тюменский государственный нефтегазовый университет**,
Центр дистанционного образования

Защита состоится 23 декабря 2011 года в 12 часов на заседании диссертационного совета К 212.274.05 по защите диссертаций на соискание ученой степени кандидата филологических наук при Тюменском государственном университете по адресу: 625000, г. Тюмень, ул. Республики, 9, ауд. 211.

С диссертацией можно ознакомиться в читальном зале ИБЦ Тюменского государственного университета по адресу: 625000, г. Тюмень, ул. Семакова, 18.

Автореферат разослан 19 ноября 2011 года.

Ученый секретарь
диссертационного совета
кандидат филологических наук,
доцент Т.В. Сотникова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Развитие глобальных сетей, а также технологий создания интеллектуальных систем обуславливает значительный интерес к исследованиям, направленным на автоматическую обработку данных, и прежде всего - к различным видам семантического анализа текста. Семантический анализ текста позволяет извлекать информацию о фактах, ключевых понятиях и их взаимосвязях, с последующим представлением материала в виде определенным образом структурированного, смыслового описания.

Понятие *смысл текста* не имеет однозначного формального определения. Мы будем использовать данный термин в трактовке И.А.Мельчука: «смысл – это инвариант всех синонимических преобразований, т.е. то общее, что имеется в равнозначных текстах» (И.А. Мельчук).

Построение смыслового описания текста может решать многие практические задачи, в том числе: семантический поиск: выявление фактов, в которых принимают участие конкретные ключевые понятия; обработка текста на естественном языке в системах управления контентом; проверка ответов учащихся в системах контроля знаний при использовании открытой формы тестирования. Такого рода задачи характеризуются необходимостью сравнения двух текстов друг с другом на смысловое соответствие с учетом *предикативных* отношений в тексте.

Как известно, тестовые задания для контроля знаний обычно составляются на основе учебного материала – текста лекции, учебного пособия, методических указаний и т.п. Для проверки необходимо производить сопоставление смыслов текста ответа и текста лекции, на основе которой было сформулировано тестовое задание. В современных системах тестирования автоматизированная проверка открытой формы реализуется при условии ввода ответа на *ограниченном* естественном языке. Для проверки ответа на *неограниченном* естественном языке необходимо предусматривать возможность использования синонимов и различных вариантов построения фразы без ограничений на членимость текста по

предложениям, что дает возможность испытуемому выразить мысль в произвольной форме.

Традиционные подходы к описанию естественного языка рассматривают текст на нескольких уровнях. Применительно к проблеме семантики текста разделение на уровни следующее: фонетический, фонологический, лексико-морфологический, синтаксический, уровень смысла текста (И.А. Мельчук). Предикативные отношения выявляются на синтаксическом уровне в рамках одного предложения. Что касается смыслового описания текста в целом, то во многих задачах (к примеру, в информационном поиске) оно реализуется с применением частотного анализа и вероятностно-статистических методов. При этом смысловая структура текста не может дать представления о предикативных отношениях между ключевыми понятиями. Следовательно, для решения перечисленных выше практических задач необходимо разработать методы и алгоритмы, использующие синтаксический анализ предложений в качестве основы для построения смыслового описания всего текста, а также сформировать количественные критерии оценки соответствия смыслов текстов.

Цель работы – разработка методов и алгоритмов определения смыслового соответствия ответа на тестовое задание контенту, по которому составлен тест.

Для достижения поставленной цели в работе решались следующие **задачи**:

1. Исследовать существующие подходы к автоматизации смыслового анализа текстов на естественном языке;
2. Формализовать описание синтаксической структуры предложений учебного контента;
3. Построить математическую модель смыслового описания контента;
4. Разработать алгоритм перехода от синтаксической структуры предложений к семантической сети контента, отображающей предикативные отношения между объектами-понятиями.
5. Сформулировать критерии оценки сходства смыслового содержания контентов и построить алгоритм сопоставления.

6. Разработать инструментальный программный комплекс для формирования тестовых заданий открытой формы на основе смыслового описания учебно-методического материала и автоматической проверки результатов тестирования.

Объект исследования: модели и алгоритмы установления смыслового соответствия контентов в системе тестирования в процессе проверки тестовых заданий открытой формы.

Предмет исследования: условия и средства получения смыслового описания учебного контента на базе синтаксической структуры предложений с учетом предикативных отношений между понятиями контента.

Методы исследования.

Лингвистические методы: *м*етод многоуровневого семантического анализа, включающий синтаксический анализ (синтаксический уровень); формальный, функциональный, категориальный анализ (лексико-морфологический уровень); метод семантических сетей (уровень текста).

Математические методы: методы теории множеств; методы теории графов; методы построения и анализа алгоритмов.

В качестве теоретических предпосылок используются:

- работы, посвященные теории «Смысл-Текст» (И.А. Мельчук, Л.Л. Иомдин, Ю.Д. Апресян, И.М. Богуславский, А.К. Жолковский)
- работы по изучению синтактико-семантических отношений в структуре предложения (Ч. Филмор, И.М. Богуславский, Н. Хомски, А.В. Гладкий и др.)
- работы по применению частотно-вероятностных методов лингвистического анализа (Г.Г. Белоногов, А.А. Хорошилов и др.)
- работы отечественных и зарубежных ученых по созданию прикладных систем автоматической обработки текста (Р.Г. Пиотровский, Н.Н. Леонтьева, В.Ш. Рубашкин, Э.В. Попов, А.Е. Ермаков, А.В. Гаврилов, Р.К. Крос, Ж.К. Гардэн, Ф. Леви, С.А. Шумский).

Материалом для исследования послужили:

- тексты лекций учебно-методических комплексов по дисциплинам «Интеллектуальные информационные системы», «Системы электронной коммерции», «Технологии мультимедиа» общим объемом 110 тыс. словоформ;

- тексты ответов на тестовые задания открытой формы, полученных в процессе итогового контроля знаний студентов 3-го и 4-го курсов специальности 080801.65 «Прикладная информатика в экономике» Тюменского госуниверситета, общим объемом 36 тыс. словоформ.

Положения, выносимые на защиту:

1. Предложена новая математическая модель представления смыслов учебного контента в виде семантической сети, узлами которой являются именные группы, обозначающие понятия, а дуги отражают предикативные отношения, характеризуемые глагольными группами. Для учета синонимии слов и выражений лексический материал контента необходимо расширить при помощи тезауруса и толково-комбинаторного словаря.

2. В основу метода построения семантической сети контента положен следующий принцип: на основе синтаксических отношений предложений выявляются а) именные группы, представляющие имена понятий контента, б) предикативные отношения, связывающие эти понятия, в) глаголы и глагольные группы, выражающие предикацию, г) отношения кореференции именных групп, выраженные в форме буквального повтора или местоименной замены слов.

3. Разработанные методы и алгоритмы позволяют произвести количественную оценку степени смыслового соответствия текстов, выраженную в двух аспектах: содержательном и структурном. Содержательный аспект отвечает за лексический состав именных групп и предикатов контентов. Структурный аспект характеризует взаимное расположение связей сравниваемых семантических сетей.

4. Сравнение смыслов контентов при автоматизированной проверке результатов тестирования в открытой форме позволяет адекватно оценить знания испытуемых, что подтверждается численными экспериментами на разработанном нами программном комплексе «Семантик Тест».

Научная новизна исследования:

1. Разработана новая математическая модель смыслового описания учебного контента, описывающая предикативные отношения между понятиями.

2. Предложен метод и построен алгоритм перехода от синтаксического описания предложений к смысловому описанию всего текста.
3. Предложен метод и разработан алгоритм количественной оценки степени смыслового соответствия двух текстов, основанный на сопоставлении их смысловых описаний, построенных с помощью математической модели;
4. Разработан программный комплекс для формирования тестовых заданий на основе смыслового описания учебного контента, а также автоматической проверки открытой формы тестирования.

Теоретическая значимость работы состоит в разработке методики создания тестирующих программных комплексов с использованием лингвистических методов анализа учебного материала.

Практическая значимость работы заключается в возможности автоматизировать процесс проверки тестовых заданий открытой формы, а также существенно упростить их разработку, что позволяет сократить временные затраты преподавателя.

Апробация работы:

Материалы диссертации докладывались на следующих конференциях и семинарах:

- Международная научная конференция «Модернизация образования в условиях глобализации», Тюмень, 2005;
- Межрегиональная научно-практическая конференция «Информационные технологии и телекоммуникации в образовании, экономике и управлении регионом», Тюмень, 2006;
- III-я Международная научно-практическая конференция «Актуальные проблемы современных наук: теория и практика», Днепропетровск, 2006;
- Всероссийская научно-техническая конференция «Приоритетные направления развития науки и технологий», Тула, 2007;
- VI-я Межвузовская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного

пространства», Тюмень, 2007;

- II-я Межрегиональная научно-практическая конференция «Информационные технологии и телекоммуникации в образовании, экономике и управлении регионом», Тюмень, 2008.
- Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений. Третья научно-практическая региональная конференция. Тюмень, 2010.
- Экономические и экологические проблемы в меняющемся мире: сборник материалов Международной научно-практической конференции, посвященной 80-летию Тюменского государственного университета. Тюмень, 2010.
- Научно – методические семинары кафедры информационных систем Тюменского государственного университета (2005 – 2011 гг.).

Разработанный тестирующий комплекс «Семантик Тест» используется в учебном процессе в Тюменском государственном университете. Имеются свидетельства о государственной регистрации программы для ЭВМ и базы данных.

Публикации. Основное содержание диссертации представлено в 18 печатных работах, из которых 2 – свидетельства о государственной регистрации программ для ЭВМ и 3 статьи, опубликованных в ведущих рецензируемых журналах.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложений. Объем диссертации составляет 119 страниц, включая 12 рисунков и 13 таблиц. В списке литературы указано 116 наименований работ российских и зарубежных авторов.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, сформулированы цели и задачи исследования, показаны научная новизна и

значимость работы.

Первая глава содержит обзор литературных источников, посвященных проблеме построения смыслового описания текста. Рассматриваются виды семантических сетей и моделей данных, используемых для смыслового представления текста, произведен сравнительный анализ инструментальных средств синтаксического и семантического анализа.

В настоящее время не существует единого подхода к построению смысловых описаний. В то же время методы предшествующего этапа - синтаксического анализа, достаточно хорошо изучены (И.А. Мельчук, А.В. Гладкий, Н. Хомски, Н.Н. Леонтьева) и имеются все предпосылки для построения алгоритма перехода от синтаксической структуры предложений к семантической сети всего контента, отображающей предикативные отношения.

Анализ программных средств показывает, что в настоящее время смысловой анализ реализован с применением вероятностно-статистических методов с вычислением разных видов релевантности (Г.Г. Белоногов, А.А. Хорошилов, А.Е. Ермаков), а также построения семантического вектора, описывающего контекст употребления одного понятия среди других. Разработка таких средств обусловлена необходимостью анализа больших объемов неструктурированной информации. Для детального анализа одного конкретного документа с установлением предикативных отношений между понятиями такие средства непригодны.

Таким образом, для программной реализации смыслового сопоставления двух контентов требуется разработать методы и алгоритмы построения и сравнения двух и более смысловых описаний контентов на основе синтаксических структур предложений. В качестве математического аппарата целесообразно использовать методы теории графов, которые нужно модифицировать и расширить методиками учета возможной синонимии и перифраз при поиске подсети в сети. Кроме того, необходимо выработать критерии оценки степени смыслового соответствия двух контентов с учетом степени синонимии слов и

выражений.

Вторая глава посвящена формализации синтаксической структуры предложения, моделированию смыслового описания контента, разработке методов и алгоритмов построения смыслового описания текста, сравнения смысловых описаний и формулированию критерия смыслового соответствия двух текстов.

Под *контентом* (от англ. *content* – «содержимое») мы понимаем собирательный термин для текстовой составляющей информационного наполнения электронного ресурса (лекции, электронной версии учебного пособия, web-страницы, тестового задания). В данном исследовании рассматривается учебный контент, используемый для формирования тестовых заданий в системе тестирования.

В качестве исходных данных выступают синтаксические структуры предложений контента. На основе синтаксических отношений между словоформами каждого предложения строится предикатно-аргументная структура, представленная в виде набора элементарных пропозиций – двухместных предикатов. Первый аргумент предиката – именная группа подлежащего, второй аргумент – именная группа дополнения или обстоятельства.

Рассмотрим текст, состоящий из предложений, каждое из которых имеет номер *s*. Представим модель синтаксической структуры предложения в виде

$$D_s=(T, B_s, C) \quad (1)$$

где D_s – модель синтаксической структуры предложения, s – номер предложения, T – множество словоформ текста, $B_s=\{b_k\}$- множество синтаксических отношений s-того предложения, C – множество типов отношений.

Каждое синтаксическое отношение b_k определяется в виде упорядоченного набора:

$$b_k=<t_i, t_j, c> \quad (2)$$

где k - номер синтаксического отношения в предложении, t_i – i-я словоформа предложения, c – тип синтаксического отношения, $c \in C$, где C={”атрибутивный”, ”актантный”, ”обстоятельственный”}.

Словоформа t_i формально представима в виде набора:

$$t_i = <l, F^{kt}> \quad (3)$$

где i – порядковый номер слова в контенте, l - лексема, $F^{kt} = <f_1, f_2, ..., f_m>$- набор грамматических характеристик, определяющих грамматическую форму словоформы t, kt – частеречная категория.

Смысловое описание s-того предложения Q^s формально представимо в виде упорядоченного набора

$$Q^s = <U, R>_s \quad (4)$$

где $U_s = \{u_1, u_2, \ldots, u_n\}$ — множество именных групп s-того предложения;

$R_s = \{r_1, r_2, \ldots, r_v\}$ – множество смысловых отношений s-того предложения.

Смысловое отношение представимо в виде:

$$r_v = <u_m, u_n, p_v> \quad (5)$$

где u_m, u_n – именные группы, связанные предикативным отношением, p_v – метка дуги, представляющая предикат, выражающий смысловое отношение между именными группами.

Именная группа представима в виде дерева синтаксических отношений словоформ с корневой вершиной, представляющей главное слово группы – имя существительное:

$$u_m = \{(l_{k1}, l_{k2}), (l_{k2}, l_{k3}), \ldots (l_k, l_n)\} \quad (6)$$

где m – номер именной группы, l_k– лексема словоформы t_i.

Алгоритм построения именной группы представлен следующей последовательностью действий:

1. Выбрать из множества синтаксических отношений B_s элемент b_{k1}, удовлетворяющий условию $b_{k1} \in \{ b_k$: c= “актантный”$\}$, вычислить его зависимую словоформу $t_{i1} = Dep(b_{k1})$ и определить ее лексему.
2. Создать из словоформы t_{i1} корневую вершину дерева именной группы u_m.
3. Выбрать из множества синтаксических отношений B_s элемент b_k, удовлетворяющий условию $b_k \in \{ b_k$: c= “ атрибутивный”$\}$, главным словом которого является словоформа $t_{i1} = Main(b_k)$. Вычислить его зависимую словоформу $t_i = Dep(b_k)$ и определить ее лексему.

4. Создать из словоформы t_i вершину дерева именной группы u_m и соединить направленной дугой с корневой вершиной t_{i1}.
5. Действия шагов 3-4 повторяются для всех элементов b_k таких, что t_{i1}=Main(b_k) и t_i=Main(b_k) до тех пор, пока будут обнаруживаться b_k.
6. Выбрать из множества синтаксических отношений B_s элемент b_{k2}, удовлетворяющий условию $b_{k2} \in \{ b_k$: c= “актантный”$\}$ и повторить для него действия шагов 1-5.

Утверждение 1. В результате работы Алгоритма 1 формируется множество именных групп U=$\{u_m\}$, где m – номер именной группы. Каждая именная группа представляет дерево, в вершинах которого находится лексема.

Алгоритм построения смыслового отношения (5) включает следующие действия:

1. Выбрать из множества синтаксических отношений B_s элемент b_{k1}, удовлетворяющий условию $b_{k1} \in \{ b_k$: c= “актантный”$\}$.
2. Выбрать из множества синтаксических отношений B_s элемент b_{k2}, удовлетворяющий условию: $b_{k2} \in \{b_k$:c=“ актантный ”, Main(b_{k1})=Main(b_{k2})= $t_i\}$.
3. Создать смысловое отношение $r_v = <u_m, u_n, p_v>$, состоящее из следующих элементов:
   a. Именная группа u_m имеет корневую вершину t_{i1}= Dep(b_{k1}), Sent(t_{i1})= “Именительный”.
   b. Именная группа u_n имеет корневую вершину t_{i2}= Dep(b_{k2}), Sent(t_{i1})≠ “Именительный”.
   c. Предикат p_v имеет корневую вершину t_i=Main(b_{k1})=Main(b_{k2}).

Утверждение 2. В результате работы алгоритма формируется множество смысловых отношений $r_v = <u_m, u_n, p_v>$, образующих ориентированный граф s-того предложения Q^s, узлами которого являются именные группы u_m , а метками дуг - предикаты p_v.

Смысловое описание всего текста формируется из смысловых описаний

отдельных предложений путем их объединения по кореферентным именным группам:

$$Q = \bigcup_s Q^s \tag{7}$$

где Q – семантический граф текста.

Кореферентность именных групп устанавливается на основании изоморфности их деревьев.

Смысловое описание контента представимо в виде:

$$\Theta = <Q, Tr, Ts, \Pi> \tag{8}$$

где

Q – семантический граф контента, Tr – тезаурус именных групп и предикатов, Ts – толково-комбинаторный словарь, $\Pi = \{\Pi_i\}$ - набор правил перифразирования.

Тезаурус именных групп и предикатов описывается в виде набора:

$$Tr = <U, P, H> \tag{9}$$

где U – множество именных групп; P – множество предикатов; H – отношение между двумя именными группами или двумя предикатами, ставящее в соответствие каждой паре (u_m, u_n) или (p_m, p_n) значение веса $a_m \in [0,1]$ и характеризующее степень синонимии соответствующих именных групп или предикатов. Вес, равный единице, означает полное синонимическое совпадение терминов, частным случаем которого является изоморфизм именных групп или предикатов.

Толково-комбинаторный словарь представляет набор, который сопоставляет слову в каноническом виде значения лексических функций, примененных к этому слову:

$$Ts = <L, LF> \tag{10}$$

где $L = \{ l_i \}$ – множество словоформ в каноническом виде, LF – множество лексических функций для словоформы l_i. Каждая лексическая функция может возвращать одно или несколько значений, также представляющих словоформы в

каноническом виде. К примеру, лексическая функция Syn(l_i) возвращает список слов, являющихся синонимами слова l_i.

Правило перифразирования Π_i сопоставляет некоторую структуру смыслового описания другим структурам, несущим тот же смысл. Структуры в правилах описываются с применением лексических функций к элементам смыслового описания:

$$\Pi_i: [u_{n1},\ldots, u_{n2}, p_{m1},\ldots, p_{m2}] \Leftrightarrow [LF_{j1}(u_{n1}),\ldots, LF_{j2}(u_{n2}), LF_{j3}(p_{m1}),\ldots, LF_{j4}(p_{m2})]$$

При рассмотрении задачи поиска с учетом описания предметной области требования к сети Q, соответствующей по смыслу запросу Q' можно сформулировать следующим образом:

1. Для именных групп u_x, u_m' смысловых описаний Q и Q' выполняются условия: $a_x > A_{пред}$, где $u_x \in Q$, $u_m' \in Q'$, $A_{пред}$ – некоторая константа, определяющая пороговое значение степени синонимии a_x, начиная с которого именные группы считаются совпадающими по смыслу.
2. Для именных групп u_m', u_n' из $r_v' = \langle u_m', u_n', p_k' \rangle \in Q'$ и $u_x, u_y \in Q$, удовлетворяющих условию 1, существует цепочка дуг, соединяющих узлы: $r = (u_x, \ldots, u_y)$.
3. Если r представляет смысловое отношение $r_v = \langle u_x, u_y, p_z \rangle$, то вес $W_v > B_{пред}$. Вес определяется по формуле

$$W_k = K_П \cdot a_i + K_Д \cdot a_j + K_{Пр} \cdot a_v \qquad (11)$$

где $K_П$, $K_Д$, $K_{Пр}$ – параметры при весовых коэффициентах именной группы-подлежащего, дополнения и предиката соответственно.

Оценка степени смыслового соответствия двух семантических графов складывается из двух факторов: содержательного и структурного.

Содержательной мерой смыслового соответствия сети запроса Q' сети текста Q будем считать величину:

$$\varepsilon(Q, Q') = 1 - \frac{1}{\sqrt{M}} \sqrt{\sum_{k=1}^{M} (W_k - 1)^2} \qquad (12)$$

где M – число смысловых отношений в сети Q ,

W_k – вес k-го смыслового отношения, вычисляемый по формуле (11).

Структурный показатель σ(Q, Q') *смыслового соответствия* сети запроса Q' и сети текста Q:

$$\sigma(Q,Q') = \frac{(\sum_{i,j=1}^{M} c_{ij} - M)}{M(M-1)} \qquad (13)$$

где c_{ij} представляет константу, значение которой определяется взаимным расположением i-той и j-той дуг семантического графа в запросе и в тексте. Значения c_{ij} могут принимать одно из значений: 0, 0.5 и 1. Следовательно, значения структурного показателя лежат в интервале [0;1].

Третья глава посвящена описанию программного комплекса «Семантик-тест», при разработке которого использован предложенный алгоритм получения смыслового описания контента, а также поиска в этом описании фрагмента, соответствующего по смыслу ответу на тестовое задание в открытой форме.

Программный комплекс состоит из следующих компонентов:

1. *Контур синтаксического анализа* производит выделение словоформ текста, определяет грамматические и синтаксические характеристики каждой словоформы, на основании которых выявляет синтаксические отношение между ними. На вход контура поступает текст, на выходе получается набор синтаксических отношений между словоформами, определенный для каждого предложения текста. Рассмотрим пример.

На вход контура поступил фрагмент лекции: *Электронная коммерция обеспечивает проведение маркетинговых мероприятий путем использования Сети. Благодаря электронной коммерции предприятия извлекают из применения Интернета прямую прибыль.* Синтаксические отношения, полученные на выходе контура, показаны в табл. 1.

Таблица 1

Синтаксическое описание предложений контента

№ пред-ложения	№ синт. отн.	Главная словоформа [номер в предложении]	Зависимая словоформа [номер в предложении]	Падеж (предлог)*	Тип синтаксич. отношения
1.	1.	коммерция [2]	электронная [1]		Атрибутивное
	2.	обеспечивает [3]	коммерция [2]	И	Актантное
	3.	обеспечивает [3]	проведение [4]	В	Актантное
	4.	проведение [4]	мероприятий [6]		Атрибутивное
	5.	мероприятий [6]	маркетинговых [5]		Атрибутивное
	6.	обеспечивает [3]	использования [13]	Т (путем)	Актантное
	7.	использования [13]	Сети [14]		Атрибутивное
2.	1.	извлекают [4]	предприятия [3]	И	Актантное
	2.	извлекают [4]	прибыль [8]	В	Актантное
	3.	прибыль [8]	прямую [7]		Атрибутивное
	4.	извлекают [4]	применения [5]	Р (из)	Актантное
	5.	применения [5]	Интернета [6]		Атрибутивное
	6.	извлекают [4]	коммерции [2]	Д (благодаря)	Актантное
	7.	коммерции [2]	электронной [1]		Атрибутивное

* Падеж и предлог указываются только для отношения типа «актантный».
И – именительный, Р – родительный, Д – дательный, В – винительный, Т – творительный

При работе модуля используется библиотека правил полного синтаксического анализа текста на русском языке «RCO Syntactic Engine» производства ООО «Гарант-Парк-Интернет».

2. *Контур семантического анализа* — это часть программного комплекса, задачей которой является представление структуры текста в виде семантической сети. На вход контура поступает набор синтаксических отношений предложений

текста, полученный контуром синтаксического анализа. На выходе контура получается описание семантической сети текста. Рассмотрим получение семантической сети из синтаксических отношений из таблицы 1.

Лексический состав именных групп, имеющих структуру деревьев, в узлах которых располагаются лексемы, представлен в таблице 2.

Таблица 2

Именные группы предложений контента

№ предл.	№ именной группы	Родительский узел дуги: лексема [номер в предл.]	Дочерний узел дуги: лексема [номер в предл.]
1.	1.	коммерция [2]	электронный [1]
	2.	проведение [4]	мероприятие [6]
		мероприятие [6]	маркетинговый [5]
	3.	использование [13]	Сеть [14]
2.	1.	коммерция [2]	электронный [1]
	2.	предприятие [3]	-
	3.	применение [5]	Интернет [6]
	4.	прибыль [8]	прямой [7]

Полученные именные группы представляют понятия, участвующие в предикативных отношениях, определенных в рамках предложений и представленных в таблице 3.

Таблица 3

Семантические сети предложений контента

№ предл.	№ связи	Именная группа - подлежащее	Именная группа - дополнение	Предикат
1.	1.	коммерция – электронный	проведение – мероприятие, мероприятие – маркетинговый	обеспечивать
	2.	коммерция – электронный	использование – Сеть	обеспечивать

2.	1.	предприятие	прибыль – прямой	извлекать
	2.	предприятие	применение – Интернет	извлекать
	3.	предприятие	коммерция – электронный	извлекать

Далее семантические сети предложений объединяются в семантическую сеть контента по кореферентным именным группам. Нумерация именных групп становится независимой от номера предложения (см. табл. 4).

Таблица 4

Именные группы контента

№ именной группы	Родительский узел дуги: лексема	Дочерний узел дуги: лексема
1.	коммерция	электронный
2.	проведение	мероприятие
	мероприятие	маркетинговый
3.	использование	Сеть
4.	предприятие	-
5.	применение	Интернет
6.	прибыль	прямой

Смысловые отношения переопределяются в соответствии с новыми идентификаторами именных групп. Структура семантической сети контента показана в таблице 5.

Таблица 5

Семантическая сеть контента

№ связи	Именная группа - подлежащее	Именная группа - дополнение	Предикат
1.	коммерция – электронный	проведение – мероприятие, мероприятие – маркетинго-вый	обеспечивать

2.	коммерция – электронный	использование - Сеть	обеспечивать
3.	предприятие	прибыль – прямой	извлекать
4.	предприятие	применение – Интернет	извлекать
5.	предприятие	коммерция – электронный	извлекать

3. *Контур тестирования* включает:

а) Интерфейс для преподавателя, позволяющий составлять тестовые задания на основе семантической сети текста лекции, формировать тесты, назначать их студентам, просматривать результаты тестирования.

б) Интерфейс для студентов, предоставляющий возможность ввода ответов на тестовые задания.

Допустим, преподаватель составил вопрос: «*Что получают предприятия благодаря электронной коммерции?*

Студент может ввести ответ в различных вариантах построения фразы, например: *«Благодаря электронной коммерции применение Интернета приносит предприятиям прямую прибыль», «Прямая прибыль извлекается предприятиями из использования Интернета благодаря электронной коммерции», «Благодаря электронной коммерции предприятия получают прямой доход из использования Интернета»*.

Рассмотрим один из вариантов ответа студента, который поступает на вход контура синтаксического анализа, затем семантического. В результате их работы будут выявлены именные группы, показанные в табл. 6.

Таблица 6

Именные группы ответа

№ именной группы	Родительский узел дуги	Дочерний узел дуги
1.	предприятие	-
2.	доход	прямой
3.	использование	Интернет
4.	коммерция	электронный

Поскольку текст ответа состоит из одного предложения, семантическая сеть имеет вид, показанный в табл. 7.

Таблица 7

Семантическая сеть ответа

№ пред. отн.	Именная группа подлежащее	Именная группа дополнение	Предикат
1.	предприятие	доход - прямой	получать
2.	предприятие	использование - Интернет	получать
3.	предприятие	коммерция - электронный	получать

Результат сопоставления смысловых структур запроса и текста представлен в таблице 8, где каждому элементу сети ответа студента поставлен в соответствие элемент сети контента лекции, из тезауруса определена степень синонимии, рассчитан вес каждого смыслового отношения по формуле (11) и определена мера смыслового соответствия по формуле (12).

Таблица 8

Сопоставление смысловых описаний ответа и лекции

№ пред. отн.	Элементы предикатив-ного отношения	Ответ студента	Контент лекции	Сте-пень сино-нимии, a_m	Вес пред. отношения, W_i
1.	Им.группа подлежащее	предприятие	предприятие	1.00	**0.97**
	Им. группа дополнение	доход – прямой	прибыль – прямой	0.98	
	Предикат	получать	извлекать	0.92	
2.	Им. группа подлежащее	предприятие	предприятие	1.00	**0.98**
	Им. группа дополнение	использование – Интернет	применение – Интернет	1.00	
	Предикат	получать	извлекать	0.92	
3.	Им. группа подлежащее	предприятие	предприятие	1.00	**0.98**
	Им. группа дополнение	коммерция - электронный	коммерция - электронный	1.00	
	Предикат	получать	Извлекать	0.92	
$K_П = 0.40$; $K_Д = 0.33$; $K_{Пр} = 0.27$			Мера смыслового соответствия **ε**		**0.976**

Использование программного комплекса позволяет повысить эффективность работы преподавателя за счет сокращения количества времени, затрачиваемого на подготовку и проверку тестов открытой формы. Кроме того, уменьшается время изучения исходных текстовых данных (книг, электронных учебников, Интернет-источников) за счет схематичного, наглядного представления обширных объемов материала.

Четвертая глава содержит результаты экспериментального исследования корректности предложенных алгоритмов путем сравнения смысловых описаний текстов, для которых в результате экспертной оценки установлено, что они имеют сходное по смыслу содержание.

Экспериментальное исследование адекватности модели смыслового описания текста производилось в рамках апробации программного комплекса «Семантик Тест». Для организации процесса тестирования были подготовлены вопросы по дисциплинам: «Интеллектуальные информационные системы», «Системы электронной коммерции», «Технологии мультимедиа». Каждый комплект тестов включал 10 заданий открытой формы по каждой из дисциплин. В тестировании приняли участие 62 студента 3 и 4 курсов специальности 080801.65 «Прикладная информатика в экономике» Тюменского госуниверситета.

В трех группах студентов было проведено тестирование при помощи системы «Семантик Тест» с последующим автоматизированным анализом результатов. Затем те же самые ответы на задания были проверены экспертами и помечены как правильные или неправильные. Далее был произведен сравнительный анализ результатов проверки на предмет совпадения или расхождения заключений о правильности каждого ответа, выданных системой и экспертом. Показано, что при уровне значимости 0,95 достоверно утверждение: вероятность ошибочного определения системой степени смыслового соответствия составляет не более 0,06%.

Поскольку система выдает заключение на основании порогового значения

содержательной меры смыслового соответствия ε_0, предлагается методика определения этого значения. Введена весовая функция E_k, принимающая дискретные значения, причем максимальные соответствуют тем значениям ε_0, при которых наибольшее число ответов оценивается одинаково (правильные или неправильные) как системой, так и экспертами.

Использование программного комплекса в учебном процессе показало эффективность его применения при подготовке и проведении тестирования открытой формы. Произведена оценка экономии времени при использовании системы «Семантик Тест» в сравнении с использованием системы без смыслового анализа (на примере системы АСТ). Время подготовки, проведения и проверки тестовых заданий сокращается на величину до 57%. Показана зависимость эффективности использования системы от количества вопросов в тесте. Программный комплекс опробован и используется в учебном процессе.

В заключении приведены основные результаты исследования и излагаются основные выводы по диссертационной работе.

Основные результаты исследования отражены в следующих публикациях:

В ведущих рецензируемых изданиях:

1. Ивашко А.Г., Бидуля Ю.В. Моделирование смыслового описания контента // Вестник ТюмГУ. - Тюмень: Изд-во ТюмГУ, 2007. - Вып.5. - С.80-86.
2. Бидуля Ю.В. Алгоритмизация смыслового описания контента // Вестник ТюмГУ. - Тюмень: Изд-во ТюмГУ, 2008. - Вып.6. - С.195-198.
3. Ивашко А.Г., Бидуля Ю.В. Алгоритмы оценки семантического соответствия контентов // Вестник ТюмГУ. - Тюмень: Изд-во ТюмГУ, 2010. - Вып.6. - С.168-173.

В прочих изданиях:

4. Бидуля Ю.В. Использование метаданных для формирования учебно-методических материалов в системах электронного обучения // Математическое и информационное моделирование: сборник научных трудов. - Тюмень: "Вектор Бук", 2005. – Вып. 7. - С. 72-77.

5. Бидуля Ю.В. Организация структуры контента в среде разработки тестовых заданий // Модернизация образования в условиях глобализации: Сборник материалов международной научной конференции, посвященной 75-летию Тюменского государственного университета. 14-15 сентября 2005 г. / Под ред. И.Е.Видт, Г.Ф.Ромашкиной. - Тюмень: Изд-во ТюмГУ, 2005. – С. 41-44.
6. Ивашко А.Г., Бидуля Ю.В. Структура семантической сети в системе генерации тестовых заданий // Матеріали III Міжнародноі науково-практичноі конференціі "Актуалні проблеми сучасних наук: теорія та практика – 2006". - Дніпропетровськ: Наука і освіта, 2006. – Т. 10. - С.66-69.
7. Бидуля Ю.В. Реферирование текста как подготовительный этап построения семантической сети // Математическое и информационное моделирование: сборник научных трудов. - Тюмень: Издательство "Вектор Бук", 2006. – Вып. 8. - С. 46-50.
8. Бидуля Ю.В. Объектный подход в описании контента // Математическое и информационное моделирование: сборник научных трудов. - Тюмень: "Вектор Бук", 2007. – Вып. 9. - С. 11-15.
9. Бидуля Ю.В. Смысловое представление материала как этап автоматической генерации тестовых заданий // Приоритетные направления развития науки и технологий: доклады Всеросс. науч.-техн. конф./ под общ. ред. чл.-корр. Российской акад. наук В.П.Мешалкина. - г. Тула: Изд-во ТулГУ, 2007. - С. 142-143.
10. Бидуля Ю.В. Представление текста в виде семантической сети // Безопасность информационного пространства VI: сборник трудов межвузовской научно-практической конференции студентов, аспирантов и молодых ученых. Тюмень, 22-23 ноября 2007 года. - Тюмень: Изд-во ТюмГУ, 2007. - С. 54-61.
11. Бидуля Ю.В. Об одном подходе к описанию контента // Информационные технологии и телекоммуникации в экономике, управлении и социальной сфере: Материалы межрегиональной научно-практической конференции 1-30 ноября 2006г. - Тюмень: Изд-во ТюмГУ, 2007. - С. 90-92.
12. Бидуля Ю.В. Разработка программного комплекса смыслового анализа

учебных материалов.- Информационные технологии и телекоммуникации в экономике, управлении и социальной сфере: Материалы II-ой межрегиональной научно-практической конференции 15 ноября - 15 декабря 2007 г. - Тюмень: Изд-во ТюмГУ, 2008. - С. 116-118.

13. Бидуля Ю.В., Ивашко А.Г. Алгоритм построения семантической сети // Математическое и информационное моделирование: сборник научных трудов. - Тюмень: "Вектор Бук", 2009. – Вып. 11. - С. 42-50.
14. Бидуля Ю.В. Информационный поиск в семантической сети контента // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений. Третья научно-практическая региональная конференция (Тюмень, ТюмГУ, Институт математики и компьютерных наук, 14-15 апреля 2010 года). - Тюмень: "Вектор Бук", 2010. - С.50-54.
15. Бидуля Ю.В.Использование смыслового анализа в системе тестирования // Экономические и экологические проблемы в меняющемся мире: сборник материалов Международной научно-практической конференции, посвященной 80-летию Тюменского государственного университета. В 2-х ч. / Отв. за выпуск В.В.Зыков, Л.С.Киселева. - Тюмень: Печатник, 2010. - Ч.1. - С.399-401.
16. Бидуля Ю.В. Учет синонимии в модели смыслового описания контента // Математическое и информационное моделирование: сборник научных трудов. - Тюмень: "Вектор Бук", 2011. – Вып. 13. - С. 42-50.
17. Бидуля Ю.В., Губина Т.И., Губин М.В. Свидетельство о государственной регистрации программы для ЭВМ №2008615239 «Система смыслового анализа материалов и контроля знаний Семантик-тест» от 31.10.2008.
18. Бидуля Ю.В., Губина Т.И., Губин М.В. Свидетельство о государственной регистрации базы данных №2009620064 «Семантик-тест» от 29.01.2009.