УДК 004.89

Ю.И. БУТЕНКО

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Украина

МЕТОДИКА МАШИННОЙ ОБРАБОТКИ ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ КЛАССА «ЯЗЫК СТАНДАРТОВ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ»

Описана методика машинной обработки документов из нормативной базы (НБ) в процессе формирования нормативного профиля для сертификации программного обеспечения систем критического применения. Рассмотрен подход к моделированию НБ программных систем с использованием основных положений теории иерархических систем. Приведена классификация документов НБ, а также результаты их грамматического анализа. Предложен подход к представлению НБ в виде онтологической системы и технология синтеза на этой основе диалоговой системы поддержки принятия решений сертификационным аудитором.

***Ключевые слова**: программное обеспечение, экспертирование программного обеспечения, нормативная база, нормативный профиль, синтаксический анализ, семантическая информация, компрессия текста, набор ключевых слов, ядро семантической целостности.*

Введение

Экспертиза программного обеспечения (ПО), являясь основным механизмом оценки его соответствия предъявляемым требованиям и нормативным профилям, в значительной мере определяет реальные возможности в обеспечении необходимого уровня безопасности и качества в отраслях и социальных сферах, связанных с системами критического применения [1-3].

В то же время экспертиза ПО является слабо формализованным и слабо структурированным видом профессиональной деятельности. Велик субъективизм и влияние человеческого фактора (опыт специалиста-сертификатора, уровень профессиональной подготовки) на результаты экспертных оценок. При этом проблематичными остаются вопросы полноты и достоверности результатов экспертизы конкретных проектов. Весьма высока трудоемкость и стоимость экспертизы ПО сложных высокотехнологичных проектов [2].

Динамика развития международной и национальной нормативной базы программной инженерии обусловливает необходимость постоянного проведения работ по формированию гармонизированных нормативных профилей, отражающих реальные потребности повышения качества продукции в конкретных прикладных областях [2]. Нормативные профили (НП) должны включать в качестве обязательных элементов требования к характеристикам качества ПО, метрики и методы их оценки. Разработка НП представляет сложно-протекающий процесс, выполняемый уполномоченными международными или национальными организациями. Данный процесс характеризуется рядом противоречий и порождаемых ими проблем:

- слабоструктурированная и слабоформализованная область деятельности, велико влияние человеческого фактора (субъективизм);
- необходимость обеспечения возможно более полного учета имеющихся на период разработки опыта и знаний в конкретных прикладных областях;
- значительная трудоемкость и продолжительность разработки НП.

Для преодоления указанных противоречий и связанных с ними проблем целесообразным является использование средств автоматизации. Однако следует заметить, при использовании стандартных средств автоматизации возникает противоречие между требуемым уровнем эффективности процессов компьютерной поддержки сертификации и тем уровнем, который могут обеспечить стандартные средства:

На этапе сбора данных проблема вызвана наличием большого объема информации, подлежащей обработке – исходя из этого, возникает необходимость проведения классификации объектов НБ. Кроме того, необходимо разработать специальный механизм отражения взаимосвязей между отдельными объектами в НБ. Стандартные средства не могут этого обеспечить из-за недостаточно развитых средств моделирования семантики.

Как видно из вышесказанного, традиционная информационная технология, основанная на работе с данными, здесь мало применима. Необходим переход к использованию интеллектуальной информационной технологии с элементами онтологического

инжиниринга, результатом реализации которой станет диалоговая информационно-аналитическая система в форме интеллектуальной системы поддержки принятия решений (ИСППР) [3].

Цель статьи состоит в изложении специальной методики машинной обработки текстов в процессе экспертирования систем с интенсивным использованием ПО, данная методика является основой для реализации интеллектуальной технологии формирования НП путем анализа текстов технической документации на экспертируемую систему.

1. Общая формулировка проблемы и постановка задачи исследования

Проблема состоит в недостаточной эффективности существующей технологии экспертирования программного обеспечения систем критического применения, в особенности на этапе формирования нормативного профиля требований к программному обеспечению систем критического применения. Указанная проблема имеет место в силу необходимости применения значительной доли рутинного труда сертификационным аудитором (СА) в процессе просмотра и отбора документов из НБ, непосредственным образом относящихся к объекту сертификации. Конструктивный путь решения проблемы состоит в разработке специальной методики синтеза в диалоговом режиме НП, а на ее основе – программного средства поддержки принятия решений сертификационным аудитором.

Постановка задачи автоматического формирования НП, на содержательном уровне имеет следующий вид:

Исходные данные: нормативная база стандартов программной инженерии разных уровней.

Необходимо получить: нормативный профиль экспертируемого ПО в виде подмножества требований из стандартов, относящихся к конкретному запросу пользователя.

2. Описание основных этапов методики машинной обработки текстов из нормативной базы

Традиционная технология создания и развертывания ЭС включает в себя ряд принципиальных недостатков, которые препятствуют эффективному использованию таких систем при сертификации программного обеспечения систем критического применения. Во-первых, пустые оболочки ЭС, на базе которых преимущественно создаются экспертные системы, в силу своей универсальности не дают возможности учесть особенности присущие сертификации вообще, и специфику сертификации программного обеспечения систем критического применения. Во-вторых, использование таких оболочек нуждается в кропотливой работе по созданию базы знаний для экспертов сертификационного центра, чья нормативная база составляет сотни документов. Указанные выше проблемы могут быть эффективно решены путем использования онтологических моделей знаний в качестве среды для построения баз знаний ЭС. Как известно онтологические модели знаний предоставляют возможности: синтеза предметно-ориентированных баз знаний, что решает первую проблему; повторное использование знаний, что решает вторую проблему.

Исходя из вышесказанного, возникает необходимость разработки специальной методики формирования нормативного профиля, включающей в себя четыре главных этапа:

1) моделирование структуры нормативной базы с использованием иерархического анализа;

2) моделирование терминосистемы по результатам частотного анализа, учитывающего семантику;

3) моделирование онтологии классов стандартов из НБ, основанных на родо-видовых связях;

4) разработка интеллектуальной диалоговой системы в виде экспертной системы, интеллектуальным ядром которой будет синтезированная на предыдущем этапе система онтологий.

3. Моделирование структуры нормативной базы с использованием иерархического анализа

Исходными данными для реализации текущего этапа методики являются стандарты, построенные в иерархическом порядке согласно их уровням: международные, отраслевые, национальные. Стандарты и нормативно-методические документы предприятий ПО относятся к документам, обладающим высокой степенью формализованности и, как следствие, внутренней иерархической структурой. При поиске информации в коллекции сложно структурированных текстов стандартов недостаточно получить лишь список релевантных документов в качестве поисковой выдачи по причине больших объемов и высокой сложности документов. Повышение эффективности поиска в таких документах может быть достигнуто, если в качестве поисковой выдачи будут получены не только документы, но и цитаты из них – точные дословные выдержки из текста, обладающие смысловой законченностью. Цитаты могут быть получены с помощью анализа иерархической структуры текстов стандартов и НД, и далее могут быть уточнены с применением синтаксического анализа. В результате может быть получена компактная поисковая выдача, в которой отсечен значительный объем информации, нерелевантный запросу.

В процессе анализа нормативной базы выявлено три типа стандартов и НД согласно их компози-

ционной структуре: документы со строгой сквозной нумерацией, документы без нумерации и документы смешанного типа.

Формирование иерархической структуры текстов стандартов и НД необходимо осуществлять в два этапа: вначале провести анализ композиционной структуры текста, а затем - анализ синтаксических структур предложений нормативной базы.

Анализ исследуемых текстов из состава НБ показал, что они отличаются композиционной стройностью и логической последовательностью изложения материала, компактностью и краткостью. Эти тексты строятся по стандартной схеме и содержат четко определенный набор элементов. Такое построение текстов стандартов и НД обусловливает возможность использования технических средств для их обработки. На рис. 1 приведен фрагмент иерархической структуры НБ.

Проведенный синтаксический анализ НБ показал следующие уровни синтаксических конструкций: уровень сложных предложений, уровень простых предложений и уровень членов предложения. В сложноподчиненных предложениях также выявлено три вида придаточных предложений: часть-целое, причина-следствие, условие-причина.

Композиционный и синтаксический анализы создают основу для реализации следующего этапа методики, а именно построения терминологической системы.

4. Моделирование терминологической системы по результатам частотного анализа, учитывающего семантику

На данном этапе формируется терминологическая система. Для этого необходимо провести лексический анализ текстов стандартов, основывающийся на частотном подходе. Среди лексем необходимо определить какие лексемы обозначают понятия. Это могут быть как имена существительные так и целые словосочетания, состоящие из нескольких слов, но обозначающие лишь одно понятие. Терминологическую систему строят на основании родовидовых отношений.

Во многих работах предполагается, что знания для создания терминосистемы будут извлекаться из терминологических словарей. Однако, в настоящее время не существует полных и адекватных терминологических словарей по сертификации и стандартизации ПО систем критического применения.

Также ныне существующие словари не могут быть использованы для построения нужной терминосистемы в силу специфики предметной области.

В силу перечисленных выше причин для построения терминосистемы следует привлечь специалиста–терминолога, который вместе с экспертом сможет построить адекватную и полную терминосистему.

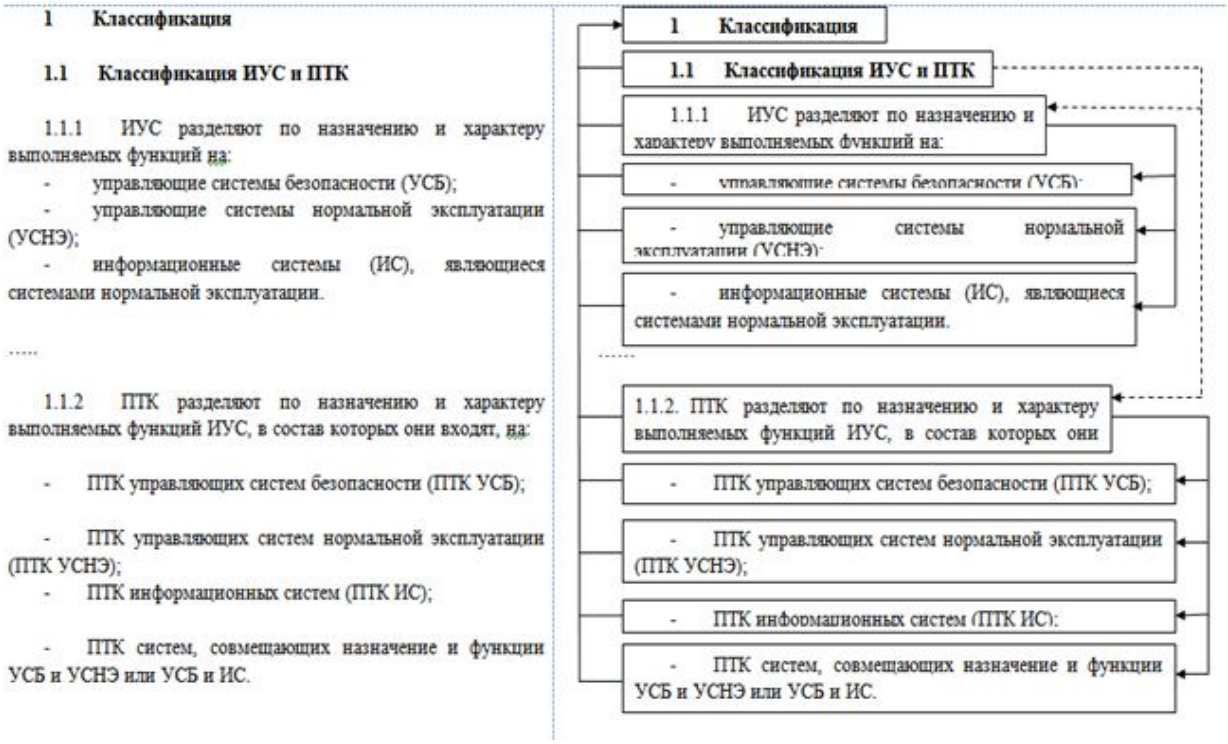

Рис. 1. Фрагмент текста стандарта и результат анализа его иерархической структуры

Структура терминосистемы должна определять связи терминов, переходы внутри общей совокупности терминов; описывать семантику, синтактику и прагматику отдельных терминов; включать описание набора семантических предикатов, регулярно связывающих термины в текстах стандартов. Для построения терминосистемы следует использовать аппарат теории семиотического моделирования, предоставляющий математический базис для построения систем такого типа. В качестве обобщенной семантической модели НБ программной инженерии квадрат Д. А. Поспелова. В этом квадрате первая вершина определяет синтаксис, или способ кодирования знака, вторая—семантику, или понятие о знаке, третья соответствует прагматике—тем процедурам, которые связаны с этим знаком, четвертая—множеству знаков, или фрагменту некоторой структуры на множестве знаков (она играет роль денотата метазнака). Фрагмент структуры на множестве знаков обладает собственным именем, выделяющим его среди остальных. Это имя представлено в вершине 1, понятие о фрагменте дано в вершине 2, а связанные с ним действия—в вершине 3. Стороны квадрата и его диагональ соответствуют различным процедурам, связывающим компоненты знака. Метазнак образует вершины 1, 2 и 3 квадрата.

В соответствии с выбранной моделью представления знака в семиотической системе словарные статьи тезауруса можно осуществить в виде системы фреймов, а затем транслировать эту систему в онтологическую среду.

В каждой словарной статье описывается одно понятие/термин. При этом в состав словарной статьи входят следующие элементы: денотат понятия, дефиниции понятия, свойства понятия, синонимы, оппозиты, список терминов, с которыми данное понятие имеет отношения. Описанный подход дает возможность построить онтологии классов стандартов из НБ, так как в терминологической системе отражены понятия, связи между ними, определены отношения между понятиями и др.

5. Моделирование онтологий классов стандартов из НБ с использованием родо-видовых отношений

Нормативную базу целесообразно представлять в виде набора онтологий классов стандартов. Каждая из этих онтологий строится по следующему алгоритму:

- выделение концептов — базовых понятий данной предметной области;
- определение «высоты дерева онтологий» – числа уровней абстракции;
- распределение концептов по уровням;
- построение связей между концептами — определение отношений и взаимодействий базовых понятий;
- консультации с различными специалистами для исключения противоречий и неточностей.

Онтологическая система для формирования нормативного профиля состоит из онтологии верхнего уровня, онтологии источника знаний, онтологий предметных областей, онтологии задач и методов, онтологии-приложения и онтологии-запроса.

Отношения между онтологиями в системе описываются общим сценарием работы системы. Каждый пользователь работает в терминах своей расширяемой онтологии запроса и посредством ее с частью онтологии приложения, сформированной для обработки данного запроса. Обработка запроса заключается в определении источников знаний, содержащих запрашиваемую информацию, ее извлечении и интеграции с последующей генерацией ответа пользователю.

Работа системы начинается с поступления на вход запроса пользователя. Затем формируется онтология-приложение для обработки данного запроса. Основная операция при формировании онтологии-приложения – операция по формированию среза. Это сложная операция, состоящая из операции выборки, соединения, отсечения и проверки на внутреннюю согласованность. Каждый сформированный под запрос пользователя срез рассматривается как самостоятельная онтология-приложение.

6. Технология синтеза интеллектуальной системы поддержки принятия решений по сертификации систем с интенсивным использованием ПО

Классическая информационная технология синтеза экспертных систем включает в себя пять этапов: идентификацию, концептуализацию, формализацию, реализацию, испытания. Недостатком данного подхода является четко выраженные границы этапов, не позволяющие модифицировать систему в процессе реализации. Данный недостаток может быть устранен применением усовершенствованной информационной технологии построения ядра экспертной системы, когда на этапе концептуализации и формализации используются методы онтологического инжиниринга.

Рассмотрим более подробно усовершенствованную информационную технологию синтеза ИСППР. На начальном этапе необходимо определить границы предметной области то есть, знание того, для какого класса программных систем будет синтезирована онтологическая модель, и насколько детальной или общей она будет. Описанный этап

определяет спектр вопросов, на которые система способна будет способна дать ответ. Необходимо учитывать, что концептуальная модель – это модель реального мира и концепты в ней должны отражать эту реальность.

На итерации формирования начальной версии онтологической модели, она оценивается и отлаживается, с использованием приложений, методов решения задач и оценок экспертов предметной области. Такой процесс итеративного проектирования имеет место в течение всего жизненного цикла разработки ИСППР.

Выводы

Описанная в статье методика машинной обработки лингвистических единиц текстов из предметной области «Сертификация систем с интенсивным использованием ПО» является основой для создания системы онтологий, содержащих знания концептуального характера о смысловой структуре нормативной базы и технической документации на системы, подвергаемые процедуре сертификации.

В дальнейшем система онтологий используется в качестве хранилища знаний ИСППР сертификационного аудитора.

Литература

1. Ахо, А. Теория синтаксического анализа, перевода и компиляции [Текст]: пер. с англ. / А. Ахо, Дж. Ульман. – М.: Изд-во «Мир», 1978. – 616 с.

2. Даниленко, В.П. Лексические требования к стандартизуемой терминологии [Текст] / В.П. Даниленко // Терминология и норма. – М.: Наука, 1972. – С. 5-32.

3. Конорев, Б.М. Концепция и принципы реализации интегрированной инструментальной системы для поддержки экспертизы и независимой верификации критического программного обеспечения [Текст] / Б.М. Конорев, В.С. Харченко, Г.Н. Чертков // Государственный комитет ядерного регулирования Украины, Государственный центр регулирования качества поставок и услуг, Сертификационный центр АСУ, Харьков, 2003. – 60 с.

Поступила в редакцию 3.06.2013, рассмотрена на редколлегии 17.06.2013

Рецензент: д-р техн. наук, проф., проф. кафедры «Программная инженерия» С.Ю. Шабанов-Кушнаренко, ХНУРЭ, Харьков.

МЕТОДИКА МАШИННОЇ ОБРОБКИ ЛІНГВІСТИЧНИХ ОДИНИЦЬ КЛАСУ «МОВА СТАНДАРТІВ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ»

Ю.І. Бутенко

Описано методику машинної обробки документів з нормативної бази (НБ) у процесі формування нормативного профілю для сертифікації програмного забезпечення систем критичного застосування. Розглянуто підхід до моделювання НБ програмних систем з використанням основних положень теорії ієрархічних систем. Наведено класифікацію документів НБ, а також результати їх граматичного аналізу. Запропоновано підхід до подання НБ у вигляді онтологічної системи та технологія синтезу на цій основі діалогової системи підтримки прийняття рішень сертифікаційних аудитором.

Ключові слова: програмне забезпечення, експертування програмного забезпечення, нормативна база, нормативний профіль, синтаксичний аналіз, семантична інформація, компресія тексту, набір ключових слів, ядро семантичної цілісності.

METHOD MACHINE PROCESSING LANGUAGE UNITS IN THE CLASS «LANGUAGE SOFTWARE STANDARDS»

I.I. Butenko

A technique for machine processing of documents from the regulatory base (RB) in the formation of normative profile for software certification systems of critical applications is described. An approach to modeling the RB of program systems with use of basic provisions of the theory of hierarchical systems is considered. Classification of documents RB, and also results of their grammatical analysis is given. Approach to RB representation in the form of ontological system and technology of synthesis on this basis of dialogue decision making support system the certified auditor is offered.

Key words: software, ekspertise of software, regulatory base, normative profile, parse, semantic information, text compression, set of keywords, kernel of semantic integrity.

Бутенко Юлия Ивановна – аспирант каф. инженерии программного обеспечения Национального аэрокосмического университета им. Н.Е. Жуковского «Харьковский авиационный институт», Харьков, Украина; e-mail*:* iuliiabutenko@yandex.ru.