

УДК 004.89

И.В. Шостак, Ю.И. Бутенко

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

ПОДХОД К АВТОМАТИЗАЦИИ ПРОЦЕССА ФОРМИРОВАНИЯ НОРМАТИВНОГО ПРОФИЛЯ ПРИ СЕРТИФИКАЦИИ ПРОГРАММНЫХ ПРОДУКТОВ

В статье описан метод компьютеризированного анализа иерархически-структурированных текстов стандартов и нормативных документов. Выявлено три типа таких документов согласно их композиционным особенностям. Указанный метод предложено реализовать в два этапа: анализ композиционной структуры текста и анализ синтаксических единиц текста. Приведен пример анализа иерархической структуры и схемы синтаксических структур в тексте стандарта.

Ключевые слова: *экспертирование ПО, нормативный профиль, иерархически-структурированный текст, композиционный анализ, синтаксический анализ.*

Введение

Современное состояние процессов компьютеризации сертификации характеризуется быстрым ростом объемов и масштабов применения специализированных комплексов, основанных на интенсивном использовании программного обеспечения (ПО), и увеличение в них доли программно реализуемых или программно поддерживаемых функций. Это обуславливает возрастание «веса» дефектов программных средств как возможных источников отказов компьютеризированных систем. Особую важность это обстоятельство приобретает для систем критического применения [3].

Экспертиза программного обеспечения, 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.

Процедура оценки ПО [3] предполагает решение ряда задач, среди которых следует выделить:

- формирование нормативного профиля (НП) – гармонизированной с международными и национальными стандартами совокупности требований, предъявляемых к данному проекту или группе проектов. НП могут быть вновь разрабатываемые государственные или отраслевые стандарты, нормативно-методические документы предприятий и общие требования спецификаций ПО;

- реинжиниринг процесса проектирования ПО и его оценка на основе НП;

- статистический анализ исходного текста, заключающийся в определении программных метрик, согласно выбранному нормативному профилю, и выполнении семантического анализа;

- динамический анализ ПО: модульное тести-

рование методом белого и черного ящиков, и интервальный анализ исполняемого модуля;

- определение степени соответствия исходного кода ПО проектной документации и НП.

В то же время экспертиза ПО является слабо формализованным и слабо структурированным видом профессиональной деятельности. Велик субъективизм и влияние человеческого фактора на результаты экспертных оценок.

Динамика развития международной и национальной нормативной базы программной инженерии обуславливает необходимость постоянного проведения работ по формированию гармонизированных нормативных профилей, отражающих реальные потребности повышения качества продукции в конкретных прикладных областях [3].

Целью статьи является описание метода компьютеризированного анализа иерархически-структурированных текстов стандартов и нормативных документов (НД) для решения одной из наиболее важных задач экспертирования ПО – формирования НП.

Постановка задачи исследования. Задача автоматизации формирования нормативного профиля на содержательном уровне может быть поставлена следующим образом: в качестве исходных данных предоставлен в электронном виде весь массив стандартов или нормативная база (НБ), а также комплект проектной документации к объекту сертификации. Результатом решения задачи является компактное подмножество документов из НБ, непосредственно относящихся к объекту сертификации.

1. Стилиевые особенности текстов стандартов и нормативных документов для автоматизации формирования нормативного профиля

Государственные и отраслевые стандарты, нормативно-методические документы предприятий и об-

шие требования спецификаций 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.

Анализ подязыка стандартов и НД показал, что его основными стилевыми чертами являются предельная ясность изложения, сжатость и лаконичность высказываний, краткость и четкость формулировок, не допускающих различных толкований, определенность содержащейся информации, динамичность и экспрессивность ее передачи, однозначность ее восприятия [4].

Сложность автоматизации задачи анализа иерархически структурированных текстов обусловлена следующими их свойствами:

- 1) как правило, разметка заголовков и маркеров (с помощью стилей, тэгов и т.д.) в документе присутствует лишь частично или отсутствует;
- 2) заголовки из разных уровней иерархии могут быть неотличимы по виду;
- 3) заголовок и ссылка на него в тексте могут иметь одинаковый вид;
- 4) разнообразие конфигураций непрерывных текстовых фрагментов: предложение может состоять из нескольких таких фрагментов, один фрагмент может включать несколько предложений, группа предложений может быть вложена в предложение в виде комментария.

В процессе анализа нормативной базы автором выявлено несколько типов стандартов и НД согласно их иерархическим особенностям, а именно:

– документы со строгой сквозной нумерацией (напр., НП 306.5.02/3.035-2000. Требования по ядерной и радиационной безопасности к информационным и управляющим системам, важным для безопасности атомных станций);

– документы без нумерации, но четкой структурой с возможностью определения заголовков, подзаголовков и др. (напр., NUREG 6303. Method for Performing Diversity and Defense-in-Depth Analyses of Reactor Protection Systems);

– смешанный тип – есть как нумерация, так и заголовки, выделенные жирным шрифтом, но не пронумерованные или нумерация по разделам – в каждом новом разделе нова нумерация (напр., NS-G-1.1. ПО для систем, базирующихся на компьютерах, важных для безопасности атомных станций. Руководство по безопасности).

Исходя из изложенного выше, предлагаемый метод к анализу иерархической структуры текстов стандартов и НД целесообразно реализовывать в два этапа.

Первый этап связан с анализом композиционной структуры текста и частности распознаванием нумерации (маркировки) разделов стандартов. Определение нумерованных (маркированных) фрагментов позволяет более точно определить границы предложений и сформировать их в виде, удобном для последующей компьютеризированной обработки, которая дает возможность на втором этапе определять синтаксическую структуру предложения с целью выявления терминологических единиц [2] и их связей для последующего построения соответствующей онтологии предметной области.

2. Метод структурной композиционной формализации нормативной базы

Структурная формализация текста – формализация самого способа изложения текстового материала. Она проявляется во всей номенклатурной организации подязыка от процессов его классификации до внешнего оформления композиционной структуры текста.

Одним из средств формализации текста выступает рубрикация, являющаяся внешним выражением его композиционной структуры. Рубрикация соответствует членению формуляра и текста стандарта или НД на составные части. Она включает графическое разделение частей, использование заголовков, подзаголовков, пунктов, подпунктов, параграфов, абзацев, нумерации и пунктуации.

Заголовки (подзаголовки) являются важным средством рубрикации и, следовательно, элементом структурной формализации. Они в предельно сжатой, краткой и лаконичной форме позволяют отразить тематику и основную идею выделенной части документа, выступая тем самым как важные единицы сообщения, передающие определенную информацию.

Рассмотрение общей структуры текста стандарта или НД и его функционально-коммуникативной направленности предусматривает знакомство с его формой. Форма стандарта или НД определяется как совокупность элементов его оформления и содержания, оцениваемых с точки зрения их состава, объема, последовательности расположения и взаимной связи.

Анализ исследуемых текстов показал, что они отличаются композиционной стройностью и логической последовательностью изложения материала, компактностью и краткостью. Они строятся по единой, стандартной, строго-установленной схеме и содержат определенный набор элементов [4]. Такое построение текстов стандартов и НД делает их более удобными для зрительного восприятия, а также упрощает их обработку и обуславливает возможность использования технических средств.

Под анализом композиционной структуры текстов стандартов и НД понимается выполнение следующих действий:

1 Классификация

1.1 Классификация ИУС и ПТК

1.1.1 ИУС разделяют по назначению и характеру выполняемых функций на:

- управляющие системы безопасности (УСБ);
- управляющие системы нормальной эксплуатации (УСНЭ);
- информационные системы (ИС), являющиеся системами нормальной эксплуатации.

1.1.2 ПТК разделяют по назначению и характеру выполняемых функций ИУС, в состав которых они входят, на:

- ПТК управляющих систем безопасности (ПТК УСБ);
- ПТК управляющих систем нормальной эксплуатации (ПТК УСНЭ);
- ПТК информационных систем (ПТК ИС);
- ПТК систем, совмещающих назначение и функции УСБ и УСНЭ или УСБ и ИС.

- 1) определение в исходном тексте элементов маркировки – заголовков;
- 2) определение вложенности групп заголовков;
- 3) определение отдельных предложений (цитат) из фрагментов текста, расположенных на разных уровнях иерархии.

На рис.1 приведен в качестве примера фрагмент текста (НП 306.5.02/3.035-2000. Требования по ядерной и радиационной безопасности к информационным и управляющим системам, важным для безопасности атомных станций) и результат анализа его иерархической структуры.



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

3. Метод синтаксического анализа нормативной базы

Проведенный синтаксический анализ текстов показал, что синтаксические структуры предложений в текстах стандартов и НД в большинстве случаев однотипны. Так, например, был проанализирован стандарт «МЭК 60880. Атомные электростанции. Системы контроля и управления, важные для безопасности. Аспекты программного обеспечения компьютерных систем, выполняющих функции категории А»; в основных разделах этого стандарта имеется 513 предложений (вводная часть и приложения не были взяты под рассмотрение), из которых 294 являются простыми предложениями, 220 – сложными, в том числе к сложным предложениям относим и перечисления, которых в тексте стандарта – 59.

Простые предложения могут быть усложнены причастными и деепричастными оборотами, кото-

рые модифицируют значение того члена предложения, к которому они относятся. Сложносочиненные предложения встречаются реже, чем сложноподчиненные. Так как сложносочиненное предложение состоит из 2 и более простых предложений, то их разбор идентичен разбору простых предложений [6].

Сама структура простых предложений в составе сложноподчиненных идентична простому предложению, но следует учитывать особенности подчиненного предложения.

В текстах стандартов выявлено три вида придаточных предложений в составе сложноподчиненных, а именно: часть-целое, причина-следствие, условие-причина.

На рис. 2 представлена обобщенная схема всех видов синтаксических конструкций, представленных в текстах стандартов, в которой можно выделить три уровня: уровень сложных предложений, уровень простых предложений и уровень членов предложения.

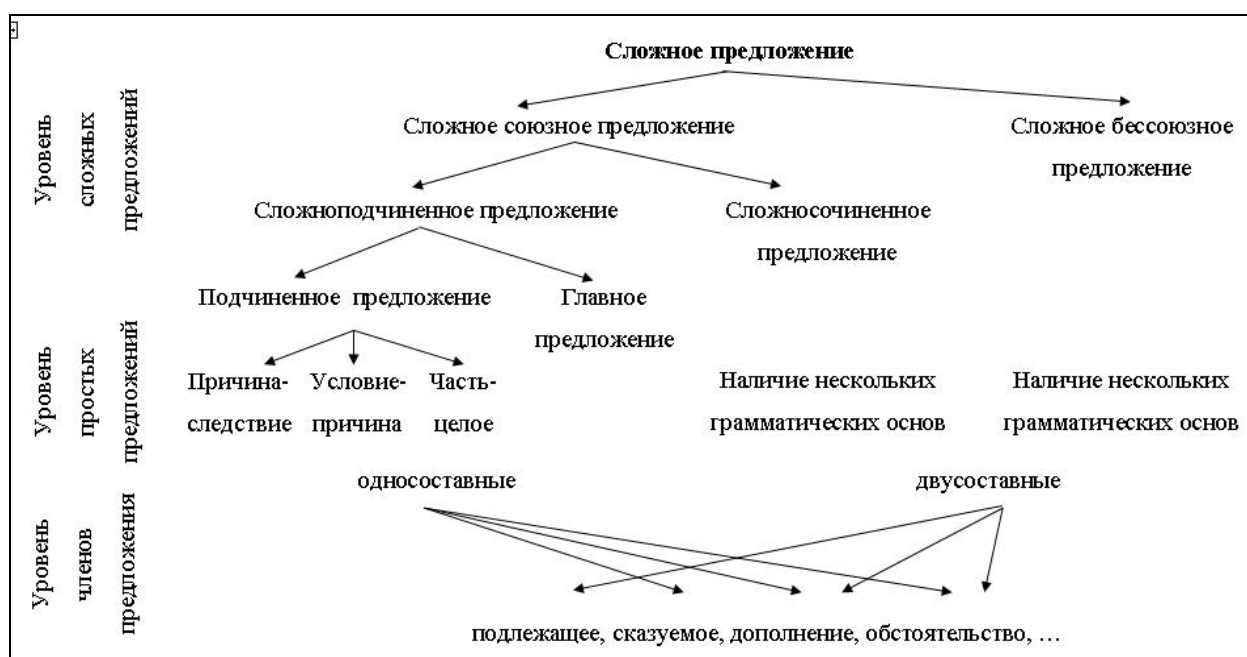


Рис. 2. Виды синтаксических структур в текстах стандартов

Алгоритмы синтаксического разбора рассмотрены и систематизированы в [1].

Проведенный синтаксический анализ нормативной базы позволяет определить терминологические единицы, а также связи существующие между такими единицами, что в свою очередь создает основу для последующего синтеза машинной процедуры, в которой эти результаты будут использованы для построения онтологической системы.

Выводы

1. В процессе анализа нормативной базы выявлено три типа стандартов и НД согласно их композиционной структуре: документы со строгой сквозной нумерацией, документы без нумерации и документы смешанного типа.

2. Описанный метод анализа иерархической структуры текстов стандартов и НД целесообразно реализовывать в два этапа: этап анализа композиционной структуры текста и этап анализа синтаксических структур предложений нормативной базы.

3. Анализ исследуемых текстов показал, что они отличаются композиционной стройностью и логической последовательностью изложения материала, компактностью и краткостью. Они строятся по единой, стандартной, строго-установленной схеме и содержат определенный набор элементов. Такое построение текстов стандартов и НД обуславливает возможность использования технических средств для их обработки.

4. 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.

5. Композиционный и синтаксический анализы создают основу для последующего синтеза машинной процедуры, в которой полученные результаты будут использованы для построения онтологической системы.

Список литературы

1. Ахо А. Теория синтаксического анализа, перевода и компиляции. Синтаксический анализ: Пер. с англ. / А. Ахо, Дж. Ульман. – М.: Мир, 1978. – 616 с.
2. Даниленко В.П. Лексические требования к стандартизуемой терминологии / В.П. Даниленко // Терминология и норма. – М.: Наука, 1972. – С. 5-32.
3. Коноров Б.М. Концепция и принципы реализации интегрированной инструментальной системы для поддержки экспертизы и независимой верификации критического программного обеспечения / Б.М. Коноров, В.С. Харченко, Н.Г. Чертков. – Государственный комитет ядерного регулирования Украины, Государственный центр регулирования качества поставок и услуг, Сертификационный центр АСУ. – Х., 2003. – 60 с.
4. Лотте Д.С. Основы построения научно-технической терминологии. Вопросы теории и методики / Д.С. Лотте. – М.: Издательство Академии Наук СССР, 1961. – 157 с.
5. Люгер Дж. Искусственный интеллект: стратегии и методы решения сложных проблем: Пер. с англ. / Дж. Люгер. – 4-е изд. – М.: Издательский дом «Вильямс», 2005. – 864 с.
6. Селіванова О. Сучасна лінгвістика: термінологічна енциклопедія / О. Селіванова. – Полтава: Довкілля-К, 2006. – 716 с.

Поступила в редколлегию 22.10.2010

Рецензент: д-р техн. наук, проф. В.Н. Рудницкий, Черкасский государственный технологический университет, Черкассы.

**ПІДХІД ДО АВТОМАТИЗАЦІЇ ПРОЦЕСУ ФОРМУВАННЯ НОРМАТИВНОГО ПРОФІЛЮ
ПІД ЧАС СЕРТИФІКАЦІЇ ПРОГРАМНИХ ПРОДУКТІВ**

І.В. Шостак, Ю.І. Бутенко

У описано метод комп'ютеризованого аналізу ієрархічно-структурованих текстів стандартів та нормативних документів. Виявлено три типи таких документів відповідно до їхніх композиційних особливостей. Вказаний метод запропоновано реалізувати у два етапи: аналіз композиційної структури тексту та аналіз синтаксичних одиниць тексту. Наведено приклад аналізу ієрархічної структури та схеми синтаксичних структур у тексті стандарту.

Ключові слова: експертування ПЗ, нормативний профіль, ієрархічно-структурований текст, композиційний аналіз, синтаксичний аналіз.

**APPROACH TO AUTOMATISATION OF NORMATIVE PROFILE FORMING FOR SOFTWARE PRODUCT
CERTIFICATION**

I.V. Shostak, I.I. Butenko

In the article it is described the method of computer analysis of hierarchically structured texts from standards and normative documents. It is found three types of such documents according to their compositional characteristics. The mentioned method is proposed to implement in two stages: analysis of compositional structure and syntaxes analysis of texts. It is given the example of hierarchical structure analysis and scheme of syntaxes structures in the text of the standard.

Keywords: software expertise, normative profile, hierarchically structured text, compositional analysis, syntaxes analysis.