OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ПРИМЕНЕНИЕ НАВИГАЦИОННОЙ СТРУКТУРЫ ЭЛЕКТРОННОГО АРХИВА ПРОЕКТНОЙ ОРГАНИЗАЦИИ В ЗАДАЧАХ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕХНИЧЕСКИХ ДОКУМЕНТОВ

Субхангулов Р.А., Филиппов А.А.

Ульяновский государственный технический университет,
г. Ульяновск, Россия

subkhangulov-ruslan@yandex.ru
al.filippov@ulstu.ru

В работе рассматривается применении интеллектуальных методов и алгоритмов анализа текстовых документов с целью построения навигационной структуры электронного архива (ЭА) проектной организации. Представление содержимого ЭА в виде иерархии кластеров, содержащих технические документы, близкие по тематике в контексте применяемых стандартов проектируемых систем, позволяет сократить пространство поиска и тем самым ускорить процедуры нахождения требуемых документов по их содержанию.
Ключевые слова: электронный архив, стандарты проектирования, онтология, навигационная структура.

Введение

Современная крупная проектная организация обладает значительным по объему ЭА конструкторской и технической документации, большая часть которой представлена в текстовом неструктурированном виде. Фактически, такой ЭА текстовой документации содержит в себе опыт и знания большого количества высококвалифицированных специалистов, которые на протяжении многих лет занимались разработкой и проектированием сложных систем. При увеличении объема ЭА затрудняется анализ документов по заранее заданным реквизитам, а от лиц, занимающихся проектированием сложных технических систем, требуются навыки в области семантической обработки большого объема технической документации, а также глубоких знаний предметной области. В результате, довольно часто важный опыт предыдущих разработок, зафиксированный в ЭА, остается невостребованным и, как следствие, увеличивается время выполнения цикла опытно-конструкторских работ.

Решение указанной проблемы может основываться на применении интеллектуальных методов и алгоритмов анализа текстовых документов с целью построения навигационной структуры ЭА технической документации. Представление содержимого такого архива в виде иерархии кластеров, содержащих технические документы, близкие по тематике в контексте применяемых стандартов проектируемых систем, позволяет сократить пространство поиска и тем самым ускорить процедуры нахождения требуемых документов по их содержанию.

Учет специфики проектных знаний приводит к необходимости формирования онтологии проектной организации особой структуры, включающей в себя особенности процесса проектирования в форме системы понятий предметной области, отношений между ними и функций интерпретации. Таким образом, ЭА должен обладать свойствами интеллектуальной системы. Актуальной является задача разработки моделей, методов и алгоритмов построения навигационной структуры ЭА технической документации на основе предметно-ориентированной кластеризации слабоформализованных информационных ресурсов.

1. Формальная модель онтологии электронного архива

1.1. Основные требования к онтологии с точки зрения процесса проектирования

Одно из определений процесса проектирования дано Норенковым И.П. в работе [Норенков, 2002]: «Проектирование технического объекта – это создание, преобразование и представление в принятой форме образа этого еще не

существующего объекта». В каждом конкретном случае такая принятая форма образа проектируемого объекта определяется на основе ряда ограничений и правил, в том числе, с использованием стандартов оформления конструкторской, программной, эксплуатационной документации и другими нормативными документами. В основе любого процесса проектирования сложной системы лежит понятие жизненного цикла (ЖЦ), который отражает различные состояния проектируемой системы, начиная с момента возникновения необходимости в данной системе и заканчивая моментом ее снятия с эксплуатации [Норенков, 2002].

При построении модели предметной области в виде прикладной онтологии для решения задач анализа технической документации необходимо сформулировать основные требования к онтологии. Такие требования должны опираться на особенности предметной области и, кроме того, на особенности тех информационных ресурсов, которые подвергаются анализу. Цель применения онтологии заключается в привлечении дополнительных знаний об окружающей среде проектируемых средств при анализе документации для сокращения времени поиска необходимых документов. Фактически, для отдельно взятого информационного ресурса из ЭА технической документации онтология задает новую систему координат, в которой кластер документов может рассматриваться как группа связанных между собой по смыслу технических документов.

На основании вышесказанного, сформулируем основные требования к прикладной онтологии:

- Онтология ЭА проектной организации должна включать в себя описания применяемых в организации моделей ЖЦ проектируемых систем.
- Структура онтологии должна основываться на использовании множества возможных серий стандартов, для каждой из которых определяется свой набор понятий и отношений между ними.
- Множество понятий онтологии должно включать в себя те понятия, которые соответствуют уже реализованным проектам, результаты которых в документальном виде зафиксированы в ЭА.

1.2. Структурно-аналитическая модель прикладной онтологии

Формально онтология ЭА состоит из двух прикладных онтологий и записывается как кортеж вида:

$$O =< O^D, O^{LC}, R_A >, \qquad (1)$$

где O^D – компонент онтологии предметной области; O^{LC} – онтология ЖЦ проектируемых систем; R_A – отношение однонаправленной ассоциации между компонентами онтологии. Рассмотрим более подробно компоненты онтологии ЭА (1).

Онтологию предметной области запишем в виде кортежа:

$$O^D =< C, W, R^D, F^D >,$$

где C – множество понятий ЭА, которое образует основу понятийного аппарата проектирования сложной системы; $W = W^S \cup W^P$ – множество терминов предметной области; W^S – множество терминов на уровне стандартов; W^P – множество терминов на уровне проектов); R^D – множество отношений:

$$R^D = \{R_G^D, R_C^D, R_A^D\},$$

где R_G^D – антисимметричное, транзитивное, нерефлексивное бинарное отношение обобщения («subclass_of»); R_C^D – бинарное транзитивное отношение композиции («part_of»); R_A^D – бинарное отношение однонаправленной ассоциации.

Множество понятий C записывается следующим образом:

$$C = (C^{S_1} \cup C^{S_2} \cup ... \cup C^{S_k}) \cup C^P,$$

где $C^{S_i}, i = \overline{1,k}$ – множество понятий предметной области, рассматриваемых в рамках i-й группы стандартов, используемых в проектной организации (например, ГОСТ 34, ГОСТ 19 и т.д.); C^P – множество понятий предметной области, извлекаемых из технической документации по реализованным проектам.

Множество интерпретирующих функций представлено в виде:

$$F^D = \{F_{WC}^D, F_{C^P C^S}^D\},$$

где $F_{WC^P}^D : \{W\} \to \{C\}$ – функция, сопоставляющая набору терминов подмножество понятий предметной области, задаваемая алгоритмически; $F_{C^P C^S}^D : \{C^P\} \to \{C^S\}$ – функция интерпретации подмножества понятий на проектном уровне онтологии, позволяющая осуществить переход на уровень понятий, определенных в стандартах.

Онтология ЖЦ как компонента кортежа (1) записывается следующим образом:

$$O^{LC} =< M^{LC}, St^{LC}, R^{LC} >,$$

где M^{LC} – множество моделей ЖЦ проектируемых систем; St^{LC} – множество стадий (этапов) ЖЦ. Отношение R^{LC} имеет вид «часть-целое (part_of)» и позволяет декомпозировать стадии ЖЦ проектируемой системы в онтологии на этапы и т. д.

1.3. Представление технического документа в контексте онтологии предметной области

Технический документ (ТД) в контексте интеллектуального ЭА будем рассматривать как информационный ресурс. Любой ТД можно рассматривать как контейнер слабоструктурированной информации: с одной стороны мы имеем дело с естественно-языковым текстом, а с другой стороны ТД имеет четко выраженную структуру, определяемую различными нормативными документами.

Если рассматривать ТД с позиций «данные-метаданные», то отдельно взятый ТД можно считать носителем данных, а онтологию ЭА – метаданными. Отношения ТД с вершинами онтологии рассматриваются применительно к двум уровням: уровень стандартов проектирования и уровень проектов.

Рассмотрим виды отношений между ТД и вершинами онтологии более детально:

- отношение «connect_to» позволяет зафиксировать в архиве принадлежность ТД множеству понятий предметной области, определенных в рамках стандартов проектирования (определяется тематика содержания документа);
- отношение «contain» между разделом ТД и термом онтологии на проектном уровне и на уровне стандартов определяет набор тех терминов, которые содержатся в разделе ТД.

Структура ТД определяется набором его разделов и подразделов, порядком их следования и вложенностью. Для отношения «contain» вес определяется как f_i^j – частота встречаемости i-го термина в j-м разделе документа. Для формирования оценок терминов будем использовать следующие правила [Наместников, 2009]:

- высокочастотные термины не являются узкими или специфическими (однако, дают большое число совпадений при сравнении терминов запроса и документа);
- низкочастотные термины вносят небольшой вклад в поиск нужных документов (поскольку редкие термины дают небольшое число совпадений);
- наилучшими конденсационными терминами являются термины не слишком редкие и не слишком частые.

Частоту появления терминов в одном ТД будем сравнивать с частотой появления тех же терминов во всем объеме документов. Если частоты терминов в анализируемом документе значительно превосходят частоту терминов по всему объему документов, делается предположение, что соответствующие термины являются ценными. Математически такая зависимость выражается следующим образом [Барсегян, 2009]:

$$f_i = tfidf_i = tf_i \cdot \log\left(\frac{N}{df(w_i)}\right),$$

где $tfidf_i$ – относительная важность термина w_i в документе; tf_i – нормализованная частота встречаемости термина w_i; N – количество всех документов; $df(w_i)$ – количество документов, содержащих термин w_i.

2. Формирование навигационной структуры электронного архива

Под терминологическим окружением понятия предметной области будем понимать множество терминов (слов) из ТД по реализованным проектам в ЭА, которые наиболее близки с данным понятием в семантическом смысле.

Определение семантического расстояния между понятием и терминами в ТД будем определять, основываясь на идее анализа семантических отношений, представленной в работе [Serrano-Guerrero и др., 2005], и заключающейся в использовании «дистанции» между словами.

В документе отношение между понятием, определенном экспертом и, одновременно являющимся термином документа, и «обычным» термином, расположенных в одном предложении, должно отличаться от отношения между понятием и термином из двух разных абзацев. Кроме того, если идея повторяется в нескольких абзацах, то она может считаться более важной, чем, если бы она была зафиксирована в одном абзаце.

После определения семантических расстояний между исследуемым понятием, для которого строится терминологическое окружение, и терминами документа необходимо определить подмножество терминов, которое наиболее тесным образом в семантическом смысле связано с понятием. При определении терминологического окружения будем использовать гипотезу λ-компактности [Загоруйко, 1999] Таким образом, используя гипотезу λ-компактности, определяется подмножество терминов, которое включается в терминологическое окружение рассматриваемого понятия.

Каждое терминологическое окружение W_k понятия $C_k^{P(S)}$ представим следующим образом:

$$\{(w_{1k}, f_{1k}),(w_{2k}, f_{2k}),\ldots,(w_{ik}, f_{ik}),\ldots,(w_{lk}, f_{lk})\},$$

где w_{ik} – i-й терм k-о понятия онтологии; l_k – общее количество термов, ассоциированных с k-м понятием; f_{ik} – нормализованный семантический вес i-о термина в терминологическом окружении k-о понятия (нормализованное семантическое

расстояние между термином и понятием в рамках одного терминологического окружения).

В основе онтологического индексирования ТД лежит следующая функция:

$$F_{oV} : ch_j^d \to oV_j^d,$$

где ch_j^d – j-ый раздел ТД d; oV_j^d – онтологическое представление j-го раздела ТДа d.

Под степенью выраженности понятия онтологии интеллектуального ЭА будем понимать степень совпадения терминологического окружения понятия с набором терминов некоторого фрагмента ТД при условии, что в терминологическое окружение включены термины, наиболее близкие в семантическом отношении с понятием.

Вычисление степеней выраженности понятий онтологии для каждого раздела ТД производится с применением аппарата нечетких соответствий [Берштейн и др., 2005]. Образом множества $\tilde{W}^d$ (множество терминов ТД d) при соответствии $\tilde{\Gamma}$ будем называть нечеткое множество $\tilde{\Gamma}(\tilde{W}^d)$ в $C^{P(S)}$, определяемое выражением:

$$\tilde{\Gamma}(\tilde{W}^d) = \{< \mu_{\Gamma(W^d)}(c^{P(S)}), c^{P(S)} > | \, c^{P(S)} \in C^{P(S)}\},$$

где $\mu_{\Gamma(W^d)}(c) = \underset{w^d \in W^d}{\vee}(\mu_{W^d}(w^d) \,\&\, \mu_O < w, c^{P(S)} >)$.

Фактически, применив функцию интерпретации онтологии $F_{WC}^D : \{W\} \to \{C\}$ на уровне проектов и стандартов, получаем первоначальное онтологическое представление каждого раздела документа в виде:

$$o\hat{V}_j^d =< ch_j, \{\hat{C}_j^P \cup \hat{C}_j^S\} >, \hat{C}_j^P \subseteq C^P, \hat{C}_j^S \subseteq C^S \,|_{St_k^{LC}},$$

где $\hat{C}_j^P$, $\hat{C}_j^S$ − первоначальные (ориентировочные) наборы понятий уровней проектов и стандартов соответственно, которые требуют уточнения.

Для генетической оптимизации ТД ЭА используем метод сравнения терминологического окружения каждого понятия в онтологии предметной области уровня проектов с анализируемым текстом [Наместников, 2012]. Минимальным фрагментом анализируемого текста является отдельное предложение, а максимальным – ТД в целом, так как в различных частях (фрагментах) документа делается акцент на разных понятиях предметной области.

Проведенные эксперименты с выделенными фрагментами ТД на основе генетической оптимизации показали, что в среднем около 30% понятий в сумме дают 70% от общей степени выраженности всех понятий фрагмента ТД. Учитывая данный факт, первоначальные наборы понятий $\hat{C}_j^P$ и $\hat{C}_j^S$ j-го раздела документа расширяются наиболее значимыми понятиями каждого фрагмента.

Заключительным шагом в формировании онтологического представления ТД является применение интерпретирующей функции $F_{C^P C^S}^D : \{C^P\} \to \{C^S\}$, которая позволяет уточнить набор понятий уровня стандартов, опираясь на найденное подмножество понятий в ТД уровня проектов онтологии.

Реализуя вышеуказанные процедуры, получаем окончательные онтологические представления для каждого j-го раздела документа в следующем виде:

$$oV_j^d =< ch_j, \{C_j^P \cup C_j^S\} >, C_j^P \subseteq C^P, C_j^S \subseteq C^S \,|_{St_k^{LC}}.$$

Для нахождения формальной меры расстояния между документами представим каждое онтологическое представление документа в качестве дерева (иерархии) понятий предметной области. Такая иерархия определяется путем нахождения минимального дерева, включающего все понятия из онтологического представления [Загоруйко, 1999].

Редакционное расстояние между иерархиями определяется на основе вычисления стоимости редакционной операции, которая находится отдельно для каждого типа семантического отношения. Итоговое редакционное расстояние между иерархиями вычисляется по следующей формуле:

$$\tau_{oV}^* = \max_i(\sum_{s=1}^{m} \varphi_{S_i}(R_G^D)_s + \sum_{l=1}^{n} \varphi_{S_i}(R_C^D)_l),$$

где $\varphi_{S_i}(R_G^D)$ – стоимость редакционной операции отношения обобщения; $\varphi_{S_i}(R_C^D)$ – стоимость редакционной операции отношения «часть-целое»; i − номер группы стандартов; s − номер добавляемого отношения обобщения; l − номер добавляемого отношения «часть-целое»; S_i – принадлежность значения редакционной операции к i-й группе стандартов.

Коэффициент нормализации T_{oV} рассчитываем исходя из всех семантических отношений обобщенной иерархии. Мера расстояния между онтологическим представлениями ТД определяется с помощью следующего выражения:

$$\left\| oV^{d_1} - oV^{d_2} \right\| = \frac{\tau_{oV}^*}{T_{oV}}.$$

Для выполнения процесса формирования навигационной структуры в виде вложенного набора кластеров технических документов необходимо решить задачу настройки весов семантических отношений между понятиями онтологии на уровне стандартов, при которых качество кластеризации, определяемое выражением:

$$F^* = \frac{\max(\overline{K}_+ + \overline{K}_-, \hat{K}_+ + \hat{K}_-)}{N} \to \min \qquad (2)$$

было бы наилучшим. В выражении (2) $\hat{K}_-$ и $\overline{K}_-$ – множества отсутствующих документов соответственно в первом и во втором кластерах; $\hat{K}_+$ и $\overline{K}_+$ – множества лишних документов соответственно в первом и во втором кластерах, N – количество документов.

3. Формальный критерий эффективности применения навигационной структуры

Будем предполагать, что искомая навигационная структура ЭА ТД образуется в процессе иерархической кластеризации документов. Особо следует отметить тот факт, что сформированная навигационная структура ЭА сохраняет свою инвариантность в пределах определенной стадии (этапа) ЖЦ проектируемой системы.

На каждом шаге происходит формирование двух новых кластеров. На нулевом шаге кластеризации электронный архив представляется в виде полного множества документов D_0 (рисунок 1). На первом шаге иерархической кластеризации получаем два множества ТД, для которых справедливо выражение: $D_1 \cup D_2 = D_0$. На втором шаге аналогичным образом формируются множества ТД $D_{11}, D_{12}, D_{21}, D_{22}$, причем $(D_{11} \cup D_{12} = D_1) \cup$ $\cup (D_{21} \cup D_{22} = D_2) = D$. Указанный процесс продолжается до тех пор, пока не сработает условие остановки иерархической кластеризации ТД и не будет построена искомая навигационная структура архива.

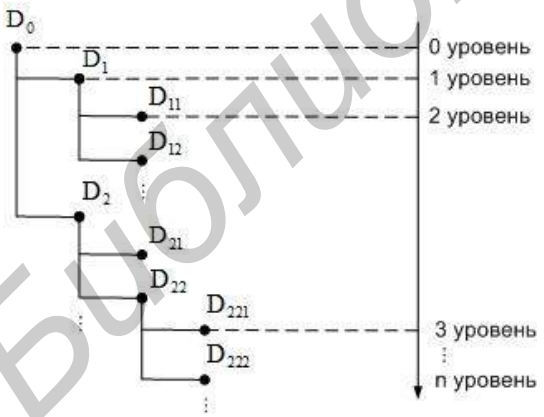

Рисунок 1 – Пример навигационной структуры ЭА

Количество документов в кластере на n-м уровне будет определяться как величина, равная $\frac{N}{2^n}$, где n – номер уровня навигационной структуры; N – количество документов в ЭА.

Применяя сформированную навигационную структуру ЭА ТД при поиске необходимого документа у проектировщика возникает преимущество, состоящее в сокращение времени нахождении документа близкого по содержанию с некоторым эталонным документом.

Пусть Δt – время, затрачиваемое на анализ одного ТД. Суммарное время анализа содержимого кластера на уровне n будет определяться величиной

$$k_n\left(\Delta t \cdot \frac{N}{2^n}\right)$$

где $k_n \in [0,1]$ – коэффициент, который определяет долю ТД, просматриваемых проектировщиком на n-м уровне. Действительно, находясь на нулевом уровне навигационной структуры, проектировщик никогда не будет анализировать все документы архива. Следовательно, k_n будет принимать наименьшее значение. При увеличении n у проектировщика появляется возможность просматривать большую долю документов в кластере. Следовательно, коэффициент k_n будет увеличиваться.

Средний процент сокращения времени поиска документов в ЭА, основываясь на навигационной структуре, определяется по следующей формуле:

$$\Delta k = \frac{k_0 \Delta t N - k_{\frac{n}{2}}\left(\Delta t \cdot \frac{N}{2^{\frac{n}{2}}}\right)}{k_0 \Delta t N} \cdot 100\% =$$

$$= \left(1 - \frac{k_{\frac{n}{2}}}{k_0} \cdot \frac{1}{2^{\frac{n}{2}}}\right) \cdot 100\%.$$

Пусть $k_0 = 0.1$; $k_{\frac{n}{2}} = 0.5$; $n = 6$, тогда получим следующее значение сокращения времени поиска документов:

$$\Delta k = \left(1 - \frac{0.5}{0.1} \cdot \frac{1}{8}\right) \cdot 100\% = 37.5\%.$$

Предполагая, что в среднем около 20% времени научно-исследовательской и опытно-конструкторской работы (НИОКР) расходуется на поиск документов в ЭА проектных организаций, можно определить, на сколько сократится время выполнения НИОКР в среднем при использовании ЭА ТД с навигационной структурой, рассматриваемой выше:

$$\Delta = \Delta k \cdot 0.2 = 7.5\%.$$

Заключение

Представление содержимого ЭА в виде иерархии кластеров, содержащих технические документы, близкие по тематике в контексте применяемых

стандартов проектируемых систем, позволяет сократить пространство поиска и тем самым ускорить процедуры нахождения требуемых документов по их содержанию.

Библиографический список

[Serrano-Guerrero и др., 2005] Serrano-Guerrero J Physical and Semantic Relations to Build Ontologies for Representing Documents / Serrano-Guerrero J., Olivas J. A., De la Mata J., Garces P. // Fuzzy logic, Soft Computing and Computational Intelligence (Eleventh International Fuzzy Systems Association World Congress IFSA), 2005, №. 1. P. 503-508.

[Барсегян, 2009] Баргесян А.А. Анализ данных и процессов: учеб. пособие. // А. А. Барсегян // Санкт-Петербург: БХВ-Петербург, 2009.

[Берштейн и др., 2005] Берштейн Л. С. Нечеткие графы и гиперграфы. / Л. С. Берштейн Л.С., А. В. Боженюк // М.: Научный мир, 2005.

[Загоруйко, 1999] Загоруйко Н. Г. Прикладные методы анализа данных и знаний /Н. Г. Загоруйко // Новосибирск: ИМ СО РАН, 1999.

[Наместников, 2009] Наместников А.М. Интеллектуальные проектные репозитории. // А. М. Наместников // Ульяновск: УлГТУ, 2009.

[Наместников, 2012] Наместников А.М. Концептуальное индексирование проектных документов на основе генетической оптимизации. // А. М. Наместников // Автоматизация процессов управления. - 2012. - №27, С. 62 – 66.

[Норенков, 2002] Норенков И. П. Основы автоматизированного проектирования: учеб. для вузов. – 4-е изд., перераб. и доп. /И. П. Норекнков // М.: Изд-во МГТУ им. Н. Э. Баумана, 2009.

APPLICATION OF NAVIGATION STRUCTURE OF DIGITAL ARCHIVE OF PROJECT ORGANIZATION IN TASKS OF THE INTELLECTUAL ANALYSIS OF CAD DOCUMENTS

Subkhangulov R.A., Filippov A.A.*

**Ulyanovsk State Technical University, Ulyanovsk, Russia*

al.filippov@ulstu.ru

This article is about application of intellectual methods and algorithms of analysis of text documents for the purpose of creation of navigation structure of digital archive (DA) of project organization. Representation of contents of DA in the form of hierarchy of the clusters containing technical documents, similar by content in the context of design standards allows reducing search space and accelerating procedures of finding of the required documents based on their contents.

Introduction

The modern project organization possesses the large DA of CAD documentation which most part is presented in the text unstructured form. The important experience of the previous development which is kept in archive remains unclaimed and, as a result, is increased the runtime of a cycle of development.

The solution of this problem can be based on application of intellectual methods and algorithms of analysis of text documents for the purpose of creation of navigation structure of DA of technical documents.

The specifics of project knowledge led to the necessity of formation the ontology of the project organization of the special structure including features of design process.

Main Part

The main requirements to application-oriented ontology:

- The ontology of DA of the project organization shall include descriptions of the life cycle (LC) models of the designed systems applied in the organization.
- The structure of ontology shall be based on use of a set of design standards, for each of which is defined the set of concepts and the relations.
- The set of concepts of ontology shall include those concepts which correspond to realize projects which are kept in a documents look in DA.

The navigation structure of DA is formed in the course of a hierarchical clustering of documents. Especially it should be noted that the created navigation structure of DA keeps the invariance within a certain stage of LC of the designed system.

On each step there is a formation of two new clusters. On a zero step of a clustering the DA is presented in the form of a full set of documents D_0 (figure 1). On the first step of a hierarchical clustering we receive two sets of documents: $D_1 \cup D_2 = D_0$. On the second step sets of documents are similarly formed $D_{11}, D_{12}, D_{21}, D_{22}$, and $(D_{11} \cup D_{12} = D_1) \cup \cup (D_{21} \cup D_{22} = D_2) = D$. The specified process proceeds until performed a condition of a stop of a hierarchical clustering of documents and the required navigation structure of archive won't be constructed.

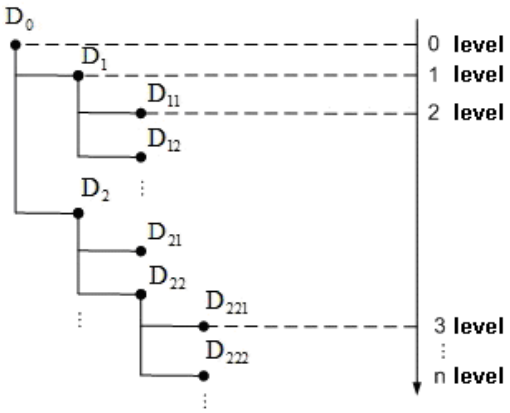

Figure 1 – Example of navigation structure of DA

Conclusion

Representation of contents of DA in the form of hierarchy of the clusters containing technical documentation, similar by content in the context of design standards allows reducing search space and accelerating procedures of finding of the required documents based on their contents.