УДК 658.51.012

А.Б. Некрасов[1], Д.Э. Лысенко[1], Н.А. Соколова[2]

[1]*Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков*
[2]*Херсонский технический университет, Херсон*

МЕТОД КЛАСТЕРИЗАЦИИ И ОЦЕНКИ МНОЖЕСТВА АНАЛОГОВ ПРОЕКТНЫХ РЕШЕНИЙ

В статье рассматриваются задачи выбора и оценки исходного множества аналогов для базы прецедентов проектных решений и списочного состава персонала по разработке образцов новой техники. Предложен метод, реализующий процедуры кластеризации множества возможных аналогов, построения рангово-классификационной шкалы сходства, определения меры сходства создаваемого образца новой техники и прецедентов-аналогов, формирования интегральной меры сходства.

Ключевые слова: *база прецедентов, множество аналогов, команда проекта, разработка новой техники, кластеризация, мера сходства.*

Введение

Обобщение и использование накопленного опыта проектирования предполагает существование системы информационной поддержки проектно-конструкторского процесса (ПКП). Основой информационного обеспечения (ИО) ПКП являются данные, которые используют проектировщики для выработки проектных решений.

Для повышения эффективности проектирования новой техники и модификации существующей целесообразно использование информации об опыте создания аналогичных проектов, а также использование тех специалистов, которые обладают таким опытом.

Одним из аспектов этого процесса является накопление в памяти стандартных, типовых, т.е. повторяющихся ситуаций, соответствующих им эффективных решений и списочного состава исполнителей, задействованных в принятии решения. Это позволяет разработчику для стандартных или близких к ним ранее встречавшихся ситуаций формировать эффективные решения, пропуская или минимизируя промежуточные работы, связанные с затратами времени, интеллектуальных и материальных ресурсов.

Постановка проблемы

Одной из главных задач разработки научно-технической продукции на основе аналогов является накопление и анализ достаточно большого и корректного для принятия решений множества прецедентов, при хранении которых используется база прецедентов.

Алгоритм формирования базы прецедентов включает следующие фазы:

1) задание весов признаков для определения уровня значимости прецедента в рассматриваемой базе;

2) кластеризация прецедентов по выявленным признакам;

3) выбор требуемого множества прецедентов на основе критерия подобия объектов.

Кластеризация базы прецедентов используется для ускорения операций выборки сходных прецедентов и предварительного разбиения базы прецедентов на компактные множества покрытия [1].

Наиболее распространенной операцией интеллектуального анализа данных является классификация. С ее помощью выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил. Однажды определенный эффективный классификатор используется для классификации новых записей в базе данных в уже существующие классы.

В качестве методов решения задачи классификации могут использоваться алгоритмы типа Lazy-Learning [2], деревья решений [3], индукция символьных правил [4], нейронные сети [5].

Существует два основных способа оценки близости прецедентов-аналогов. Первый – статистический, где для отбора прецедентов используется байесовская сеть [6]. Второй способ – введение классов эквивалентности на множестве прецедентов [7]. В статье предложены методы, решающие задачу кластеризации множества аналогов и оценки полученных классов эквивалентности.

Целью анализа является определение возможности и степени использования аналогов для проектных решений по разработке нового образца.

Решение проблемы

Предложенный метод кластеризации и оценки аналогов предполагает выполнение следующих основных этапов:

– кластеризация множества возможных аналогов проектных решений;

– построение рангово-классификационной шкалы сходства на множестве прецедентов-аналогов;

– определение меры сходства создаваемого образца новой техники и прецедентов-аналогов;

– формирование интегральной меры сходства образца и аналогов.

Кластеризация исходного множества аналогов

Кластеризация производится путем сравнительного анализа разрабатываемого образца и аналогов на основе расчета меры близости по основным элементам семантической сети технического задания. В качестве исходных данных используются аналоги из базы прецедентов, представленные в виде формализованного описания (семантической сети) технического задания на разрабатываемый образец. При этом аналогов может быть несколько, один или не существовать вообще. Предлагаемый метод наиболее эффективен в случае значительного множества аналогов, но может применяться и для двух-трех.

В результате неполноты и информационной недостаточности описания аналогов, не все они могут быть использованы для оценки требуемых характеристик разрабатываемого образца, а некоторые могут быть использованы ограниченно. Исходное множество аналогов классифицируется на пять категорий по степени информационной (Cis) и документальной обеспеченности (Cdoc) описания следующим образом (табл. 1):

Таблица 1

Классификация возможных аналогов по степени документальной и информационной обеспеченности

Класс аналога		Документальная обеспеченность		
		$0 < Cdoc < C_{doc}^{НОРМ}$	$C_{doc}^{НОРМ} \le Cdoc \le 1$	$Cdoc > 1$
Информационная обеспеченность	$Cis<0,5$	5	4	3
	$0,5 \le Cis<0,8$	4	3	2
	$0,8 \le Cis<1$	3	2	1

1. информационная и документальная обеспеченность максимальны;

2. один из коэффициентов информационной или документальной обеспеченности имеет среднее значение, другой – максимальное;

3. оба коэффициента имеют среднее значение, или один – максимальное, другой – минимальное;

4. один из коэффициентов информационной или документальной обеспеченности имеет среднее значение, другой – минимальное;

5. информационная и документальная обеспеченность минимальны.

Полученная классификация с помощью функций принадлежности переводится в лигвистическую шкалу, на основании которой может быть сделан один из следующих выводов:

- данный аналог может быть использован в полной мере для оценки реализуемости требований разрабатываемого образца (1-й класс и частично 2-й);
- аналог может быть использован частично (3-й класс, частично 2-й и 4-й классы);
- аналог не может быть использован из-за недостаточности его описания (5-й класс и частично 4-й).

Далее производится процедура определения весовых коэффициентов элементов семантической сети. Для оценки аналогов используются наиболее значимые элементы технического задания (ТЗ) на разработку образца и аналогичных проектов, например, «назначение» и «технические требования» с указанием их значений.

При кластеризации основным показателем меры сходства является степень соответствия назначения и основных технических требований (ТТ) образца и аналога. Полученная мера затем используется для корректировки близости по элементам описания эскизного проекта (ЭП). Так, «расстояние», вычисленное по элементам ЭП, увеличивается, если «расстояние» по ТЗ значительное. Качественная оценка аналогов по ТЗ предполагает разбиение их на три кластера:

а) близкие – если элементы ТЗ образца и аналога почти совпадают;

б) средние – если элементы ТЗ совпадают частично;

в) далекие – элементы ТЗ не совпадают, но достаточно близки, чтобы считать рассматриваемые образцы аналогами.

Полученное разбиение используется в последующем одним из способов:

1. Дальнейшее вычисление «расстояния» по элементам эскизного проекта внутри полученных кластеров. При этом производится ступенчатое упорядочивание аналогов: сначала по элементам ТЗ, затем по элементам ЭП.

2. Качественная оценка является основой для количественной оценки близости. При этом используется аппарат нечеткой кластеризации – мерой близости объекта к центу кластера являются значения функций принадлежности. Полученные значения затем используются для корректировки близости по элементам ЭП.

3. Если количество аналогов значительное (порядка 7-10), то достаточно рассматривать только наиболее близкие по элементам ТЗ (из первого кластера), так как оценка близости аналогов предполагает использование значительного объема информации и при большом их количестве является трудоемкой.

Построение рангово-классификационной шкалы сходства

В описании эскизного проекта (ЭП) определено больше исходных данных о параметрах создаваемого образца новой техники, чем в описании технического задания. Поэтому необходимо проводить более тщательный анализ множества аналогов, тем самым повышая достоверность результатов проектирования на основе прецедентного подхода. С этой целью на основе семантической сети эскизного проекта производится упорядочение (по мере сходства) и классификация (по типу сходства) аналогов, которые по результатам предварительного анализа технического задания попали в группы «близкий аналог» и «может рассматриваться в качестве аналога».

Установление сходства осуществляется после-

довательно по четырем основным показателям: функциональное сходство, структурное или конструктивное сходство, параметрическое, технологическое сходство. Метод предполагает сравнение элементов семантической сети ЭП разрабатываемого образца и аналога последовательно по предложенным показателям, в результате определяется преобладающий тип сходства и его степень. Для упорядочения аналогов используется метод Электра.

Этапы метода следующие.

- определение шкал показателей;
- вычисление весов показателей, определение реперов по каждой шкале;
- оценка предполагаемых аналогов по указанным показателям;
- построение графов предпочтительности по каждому показателю;
- расчет матриц согласия и несогласия;
- построение обобщенного графа;
- выделение множества аналогов и определение преобладающего типа сходства для каждого из отобранных аналогов.

В зависимости от специфики этапа разработки нового образца для дальнейшего анализа используются элементы-аналоги с соответствующим типом сходства.

Определение мер сходства осуществляется путем построения рангово-классификационной шкалы на основе таксонометрического подхода.

Метод оценки сходства аналогов и образца включает следующие этапы.

1. Осуществляется ранговое шкалирование аналогов по каждому оценочному показателю

$$R \to \{\alpha(\mu)\},$$

где α – совокупность ранговых оценок по показателям μ.

2. Вычисляется сумма рангов, полученных по каждому аналогу на множестве показателей

$$\left\{\left(\sum_{j=1}^{m}\alpha_i\right)_1,...,\left(\sum_{j=1}^{m}\alpha_i\right)_k\right\},\left(\sum_{j=1}^{m}\alpha_i\right)_1=\alpha_{\sum 1}.$$

3. Осуществляется вторичное ранжирование прецедентов-аналогов по суммарным рангам

$$\{\alpha_{\sum 1},...,\alpha_{\sum k}\}\to\alpha'(R).$$

4. Происходит разбиение аналогов на два класса на основе применения линейных ранговых дискриминаторов

$$\xi=\kappa+1/2;P_{t_j}=\left\lceil\alpha_{t_j}<\xi\right\rceil,$$

где k – число аналогов; α_{t_j} - ранг t-го аналога по j-му показателю.

5. Матрицы рангов переводятся в булеву матрицу на основе предиката P_t, где «1» получают ранги, превысившие 50% максимального ранга

$$\max_{1\le t\le K}\alpha(\mu_{jt})=k.$$

6. Выполняется операционное свертывание предикатов P_t по оценочным показателям

$$\overline{P}_{t_j}=\sum_{j=1}^{m_1}P_{t_j},$$

где m_1 – число оценочных показателей. Свертка определяет долю от всего множества оценочных показателей сходства t-го аналога, ранги которых больше $(k+1)/2$.

7. Производится дискриминация множества аналогов на два класса K_1 и K_2 в соответствии с линейным дескриптором

$$\beta=m_1-1/2;P_t=\lceil P_t>\beta\rceil.$$

Дескриптор делит аналоги на две группы – аналоги, для которых доля «хороших» оценочных показателей больше 45%, и аналоги, для которых эта доля меньше.

8. Формируются категории сходства:

$$\text{Kat.1}:\vec{\alpha}_1=\min_{K_1}\vec{\alpha}_j;\text{Kat.2}:\vec{\alpha}_2=\sum_{j=1}^{k_1}\alpha_j/k;$$

$$\text{Kat.3}:\vec{\alpha}_3=\sum_{j=k_1+1}^{k}\alpha_j/(k-k_1);\text{Kat.4}:\vec{\alpha}_4=\max_{K_2}\vec{\alpha}_i.$$

1 – лучшие ранги в первом ранговом классе K_1;

2 – средние ранги в K_1;

3 – средние ранги во втором ранговом классе K_2;

4 – худшие ранги в K_2.

Полученные категории образуют рангово-классификационную шкалу сходства на множестве выбранных аналогов.

Определение меры сходства создаваемого образца новой техники и прецедентов-аналогов

Метод позволяет оценить сходство образца и аналога четырех основных типов:

– по общефункциональным требованиям (ОФТ), что позволит судить о функциональном сходстве;

– по конструктивным особенностям, на основе которого образуется оценка структурного сходства;

– по техническим характеристикам, на основе которых делается вывод о параметрическом сходстве;

– по технологическим особенностям производства, что позволяет получить оценку технологического типа сходства.

При оценке меры *сходства по функциональным* требованиям возможны следующие четыре варианта соотношения требований:

– у аналога общефункциональные требования «шире», чем у образца. В этом случае аналог имеет, кроме совпадающих с образцом требований, ряд новых, которых нет у образца. Такая ситуация условно называется «образец – часть аналога»;

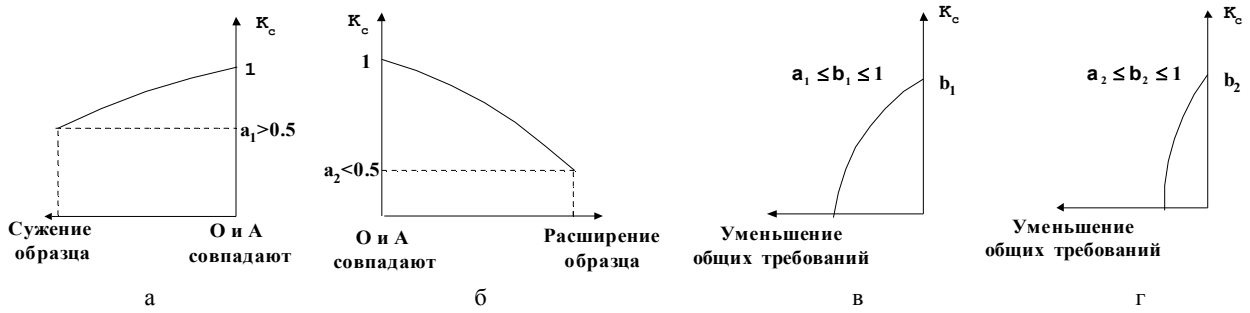
– у образца общефункциональные требования «шире», чем у аналога. В этом случае образец имеет, кроме совпадающих с аналогом требований, ряд новых, которых нет у аналога. Такая ситуация условно называется «аналог – часть образца»;

– у образца и аналога кроме общих общефункциональных требований имеются свои не совпадающие требования. Но при этом у аналога список требований «шире». Такая ситуация условно называется «аналог больше образца»;

– этот случай близок к третьему, но у образца список требований «шире». Такая ситуация условно называется «образец больше аналога».

Введем понятие «несоответствия» между целевыми требованиями к образцу и аналогу – $K_{нес}$. Очевидно, что «соответствие» есть $K_c = 1 - K_{нес}$, если использовать универсальную шкалу [0…1].

Представим возможный вид функций принадлежности для K_c в зависимости от класса рассмотренных сопоставлений образца (О) и аналога (А) по общефункциональным требованиям (рис. 1).



а б в г

Рис. 1. Графики функции принадлежности при различных вариантах ситуаций

Анализируя первый и второй варианты соотношения требований, отметим, что полное соответствие будет в случае полного совпадения множества общефункциональных требований к образцу и аналогу. Это означает единичное значение K_c для функций принадлежности. При «сужении» требований образца по отношению к аналогу происходит уменьшение соответствия K_c и соответственно увеличение несоответствия $K_{нес}$. Из-за того, что уже аналог существует, получение образца с меньшим перечнем целевых требований осуществить легче, чем для других классов, поэтому граничное значение соответствия a_1 будет не менее 0,5 (рис. 1, а).

При «расширении» требований образца по отношению к существующему аналогу происходит более резкое изменение K_c, т.к. добиться получения образца с более широкими требованиями сложнее, и поэтому граничное значение соответствия $a_2 > 0{,}5$, что указывает на резкое увеличение несоответствия (рис. 1, б).

При уменьшении общего количества требований, можно прийти к ситуации, когда аналог перестает быть аналогом (рис. 1, в) и $K_c = 0$. Очевидно, максимизация соответствия связана с ситуацией, когда образец становится частью аналога и при этом $K_c = b_1$, где $a_1 \le b_1 \le 1$.

Для четвертого варианта функция принадлежности K_c изменяется по сравнению с третьим вариантом значительно «резче», так как аналог имеет меньший спектр целевых требований, чем образец. Поэтому $a_1 \le b_1 \le 1$ (рис. 1, г).

Вид функций принадлежности, а также значения a_1, a_2 настраиваются пользователем исходя из опыта, а также класса ситуации принятия решений.

Определим средние значения K_{c_0} и $K_{нес_0}$ для каждого варианта:

1. $K_{c_0} = \frac{1+a_1}{2}$, $K_{нес_0} = 1 - \frac{1+a_1}{2}$;
2. $K_{c_0} = \frac{1+a_2}{2}$, $K_{нес_0} = 1 - \frac{1+a_2}{2}$;
3. $K_{c_0} = \frac{b_1}{2}$, $K_{нес_0} = 1 - \frac{b_1}{2}$;
4. $K_{c_0} = \frac{b_2}{2}$, $K_{нес_0} = 1 - \frac{b_2}{2}$.

Полученные средние значения могут выступать в виде «весов» каждого варианта и использоваться для усредненной оценки несоответствия функциональных требований образца и аналога.

Алгоритм метода состоит в следующем:

– определение варианта сопоставления образца и аналога с использованием визуального представления;

– получение оценки несоответствия с помощью расширения, сужения, сдвижки образца и использованием соответствующей функции принадлежности;

– расчет коэффициента несоответствия по общефункциональному назначению.

Процедура расчета параметрической меры сходства по техническим характеристикам (ТХ)

Процедура состоит из следующей последовательности шагов:

1) Сопоставление каждой характеристики образца со всеми характеристиками аналога. При этом формируется матрица сходства.

2) Анализ матрицы сходства. При этом рассматриваются варианты:

а) однозначное соответствие характеристики образца и аналога;

б) характеристика аналога не имеет соответствия у образца, эта характеристика аналога не учитывается в дальнейшем;

в) несколько характеристик образца соответствуют одной характеристике аналога. Эти характеристики объединяются с суммированием значимостей

$$W_p = \sum_i W_{ip} .$$

Мера несоответствия – расстояние d_T^i определяется экспертными методами;

г) несколько характеристик аналога соответствуют одной характеристике образца (характеристики объединяются);

д) характеристика образца не имеет соответствия у аналога. Если эта характеристика не будет учитываться в дальнейшем, то в матрицу несоответствий заносится единица. В противном случае характеристики аналога доопределяются с помощью экспертов, и анализ производится заново.

3) Вычисление степени несоответствия по характеристикам путем суммирования несоответствий из матрицы с учетом значимостей характеристик образца:

$$d_T = \sum_{i=1}^{n} W_i d_T^i .$$

4) Получение обобщенной оценки. При этом возможны варианты:

а) значения расстояний по ОФТ и ТХ близки:

$$\left| \frac{d_o - d_T}{d_{max}} \right| \le 0,1 ,$$

обобщенная оценка

$$d = \frac{d_o + d_T}{2} ;$$

б) расстояние по ОФТ ощутимо меньше, чем по ТХ:

$$\frac{d_T - d_o}{d_T} > 0,1$$

или наоборот; экспертами вводится коэффициент, отражающий несоответствие оценок:

$$d = d_o + k_1 d_T .$$

Заключение

Таким образом, в результате кластеризации исходного множества возможных аналогов получаем множество аналогов-прецедентов, которые заносятся в базу прецедентов разрабатываемого изделия новой техники. При этом указывается преобладающий тип сходства: функциональный, конструктивный, параметрический или технологический. В дальнейшем осуществляется ранжирование выбранных аналогов по степени сходства на основании оценивания выделенных признаков. Информация о наиболее близких аналогах затем используется на соответствующих этапах проектирования в процессах подготовки технической документации (конструкторской и технологической).

Список литературы

1. Schank R. Dynamic Memory: A Theory of Learning in Computers and People. – New York: Cambridge University Press. – 1982. – 205p.

2. Wang H., Dubitzky W., Dintsch I., Bell D.A. A Lattice Machine Approach to Automated Case Base Design: Marrying Lazy and Eager Learning // Proc. 17th Int. Joint Conference on Artificial Intelligence (IJCAI-99), Sweden, 1999. – P. 171-176.

3. Гупал А.М., Пономарев А.А., Цветков А.М. Об одном методе индуктивного вывода с подрезанием деревьев решений // Кибернетика и системный анализ. – 1993. – № 5. – С. 174-178.

4. Parsaye K. Rules are Much More than Decision Trees // The Journal of Data Warehousing. – 1997. – № 1. – P. 19-24.

5. Уоссермен Ф. Нейрокомпьютерная техника: Теория и практика. – М.: Мир, 1992. – 240 с.

6. Brand E., Gerritsen R. Naive-Bayes and Nearest Neighbor // DBMS. – 1998. – № 7. – P. 139-144.

7. Варшавский П.Р. Применение метода аналогий в рассуждении на основе прецедентов для интеллектуальных систем поддержки принятия решений // Тр. девятой нац. конф. по искусственному интеллекту с междунар. участием КИИ-2004. В 3-х т. Т.1. – М.: ФизМатЛит, 2004. – С. 218–226.

Поступила в редколлегию 1.08.2008

Рецензент: д-р техн. наук, проф. И.В. Чумаченко, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

МЕТОД КЛАСТЕРИЗАЦІЇ І ОЦІНКИ МНОЖИНИ АНАЛОГІВ ПРОЕКТНИХ РІШЕНЬ

О.Б. Некрасов, Д.Е. Лисенко, Н.О. Соколова

У статті розглядаються завдання вибору і оцінки початкової безлічі аналогів для бази прецедентів проектних рішень і облікового складу персоналу по розробці зразків нової техніки. Запропонований метод, що реалізовує процедури кластеризації множини можливих аналогів, побудови рангово-класифікаційної шкали подібності, визначення міри подібності створюваного зразка нової техніки і прецедентів-аналогів, формування інтегральної міри подібності.

Ключові слова: *база прецедентів, множина аналогів, розробка нової техніки, кластеризація, міра подібності.*

METHOD OF CLUSTERIZATION AND ESTIMATION OF SET OF ANALOGUES OF PROJECT DECISIONS

A.B. Nekrasov, D.E. Lysenko, N.A. Sokolova

In the article the tasks of choice and estimation of initial set of analogues are examined for the base of precedents of project decisions and list composition of personnel on development of standards of new equipment . A method, realizing procedures of clusterization of great number of possible analogues, constructions of grade-classification scale of affinity, determination of measure of affinity of the created standard of new equipment and precedents-analogues, forming of integral measure of affinity, is offered.

Keywords: *base of precedents, set of analogues, development of new equipment, clusterization, measure of affinity.*