

Ю. Ю. Якунин, А. А. Городилов

АЛГОРИТМЫ РАСЧЕТА КОМПЛЕКСНЫХ ПОКАЗАТЕЛЕЙ В ДИНАМИЧЕСКИХ СТРУКТУРАХ ПРЕДСТАВЛЕНИЯ ДАННЫХ

Описываются алгоритмы расчета комплексных показателей на множестве фактографических данных, представленных в динамических структурах с применением теории графов.

Ключевые слова: динамические структуры данных, трехдольные графы, алгоритм обхода графа, комплексные показатели.

Проблема разрыва между научными методами представления (описания) объектов реального мира и хранения информации о них в информационных системах существует уже давно и в настоящее время не нашла своего удовлетворительного решения. Так, самые передовые на сегодняшний день системы управления базами данных (СУБД), позволяющие сохранять информацию в виде объектов [1; 2] (объектно ориентированный подход [3]) или в виде глобалов [1] (иерархическое представление информации в виде дерева) – не могут охватить и доли всего многообразия представления информации современными научными методами [4]. Такой разрыв в значительной степени тормозит развитие прикладных наук и инженерии в области информационных технологий.

В связи с этим одним из существенных ограничений создания информационных систем является общепринятый способ хранения данных на базе статических структур (т. е. для описания объектов предметной области с целью хранения информации о них в базе данных заранее создается структура хранения данных). Это приводит к тому, что такая структура должна быть создана проектировщиком информационной системы еще на этапе ее проектирования и она (эта структура) не может изменяться в течение создания этой системы и ее эксплуатации. Стоит ли говорить, к каким затратам приводят изменения в структуре хранения информации на этапе создания информационной системы и ее эксплуатации... Очевидно, что если такие изменения возможны, то только в незначительной части этой структуры, иначе стоимость переработки может совпадать со стоимостью создания самой информационной системы и даже превышать её.

Любые попытки создания динамических структур хранения данных приводят к появлению задачи трансформации этих данных в один из существующих способов их хранения в СУБД (реляционный, объектный или иерархический). Ни один из перечисленных способов не способен работать напрямую с динамическими структурами. Зачастую такие задачи решаются индивидуально по мере их возникновения. В результате отсутствуют какие-либо общие подходы для решения этих задач, а реализованные в некоторых программных продуктах в виде отдельных модулей подобные задачи, представляют собой коммерческую тайну. Таким образом, существует научная задача описания динамических структур данных, обработки информации в этих структурах и хранение её в СУБД, а также частная задача расчета комплексных показателей на множестве фактографических данных в динамической структуре, решение которой и предлагается в данной статье.

Представление фактографических данных на трехдольном графе. Для описания динамической структуры данных предлагается выделять следующие категории информации [5]:

- показатели – количественные характеристики объектов;
- классификаторы – структурообразующие данные, состоящие из взаимосвязанных классов и их экземпляров (объектов);
- фактографические данные – значения показателей в отношении одного или нескольких классификаторов.

Для представления каждой категории информации будем использовать соответствующее графовое описание.

Граф показателей состоит из двух частей, одна из которых представляет собой дерево (рис. 1), содержащее вершины-категории (классификация показателей), а другая – простой граф с произвольным числом долей этого графа, вершины которого представляют собой показатели.

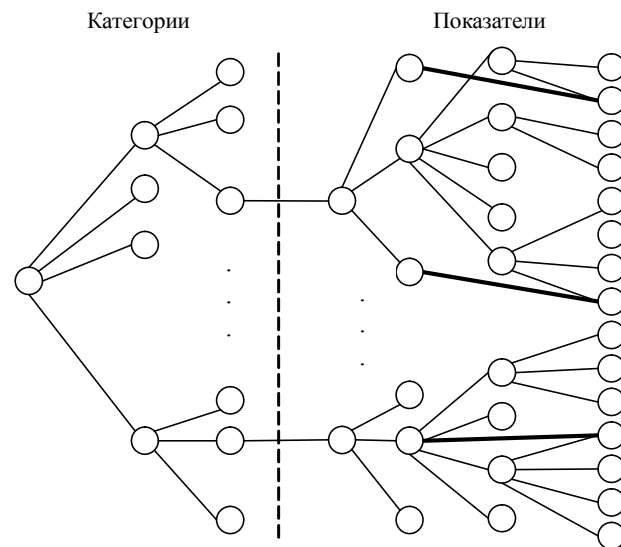


Рис. 1. Граф показателей

В части категорий ребра показывают (слева направо) разбиение категорий на подкатегории, а в части показателей ребра показывают (слева направо) разбиение агрегированных показателей на более частные до элементарных показателей (неделимых). Агрегированные показатели могут состоять не только из показателей предыдущего уровня (k), но и из показателей следующего уровня ($k + 1$), через уровень ($k + 2$) и т. д.

Рассмотрим граф классификаторов (рис. 2). Граф делится на три уровня, каждый из которых несет собственную смысловую нагрузку. Так, уровень категорий пред-

назначен для разбиения классификаторов на укрупненные категории, т. е. описания структуры классификационного дерева. На уровне классификаторов представлено множество вершин (классификаторов) с ребрами между уровнем категорий, показывающими принадлежность каждого классификатора к какой-либо категории, и ребрами между вершинами классификаторов, показывающих связи между ними, а сама вершина характеризуется не только ее наименованием, но и набором атрибутов.

Уровень значений классификаторов аналогично предыдущему уровню содержит множество вершин, каждая из которых имеет одно обязательное ребро, показывающее принадлежность значения к какому-либо классификатору, и несколько необязательных ребер, показывающих зависимость между значениями разных классификаторов. Вершины данного уровня также характеризуются наименованием и значениями атрибутов, соответствующих вершинам в уровне классификаторов, с которыми они связаны ребрами.

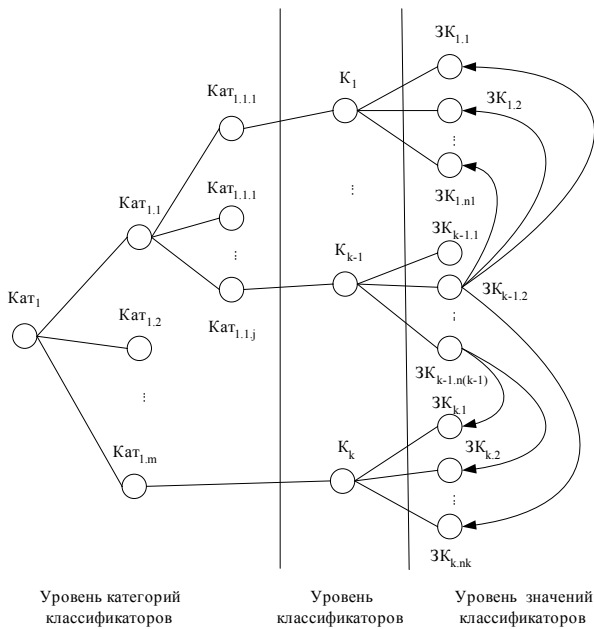


Рис. 2. Граф классификаторов:
 Кат – категория классификаторов; К – классификатор;
 ЗК – значение классификатора

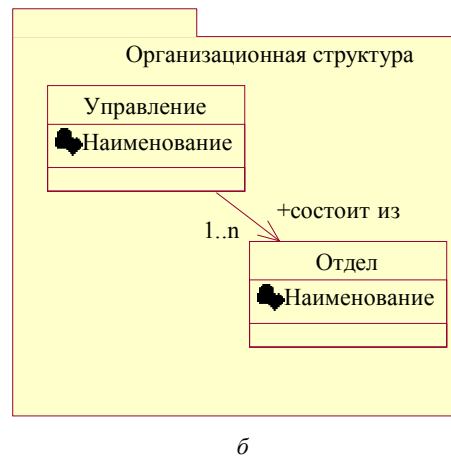
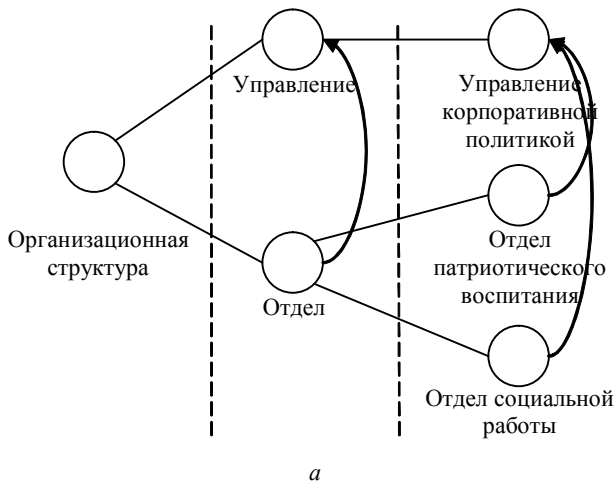


Рис. 3. Примеры представления структур: а – графовый вид; б – объектный вид

Рассмотрим в качестве примера организационную структуру некоторой организации (рис. 3) и выделим две сущности в этом примере – управление и отдел. Проведем аналогию графового представления структуры классификаторов (рис. 3, а) с другими способами представления данных, например, с объектным (рис. 3, б).

Фактографические данные представляют собой граф со множеством вершин, не имеющих ребер между собой, но имеющих ребра с вершинами графов классификаторов и показателей (рис. 4). Вершины фактографических данных представляют собой числа, которые являются количественными характеристиками одного или нескольких показателей в отношении с одним или несколькими классификаторами. Например, для показателя «зарплата» и значения классификатора «отдел социальной работы», значение фактографического данного может равняться X рублей в год.

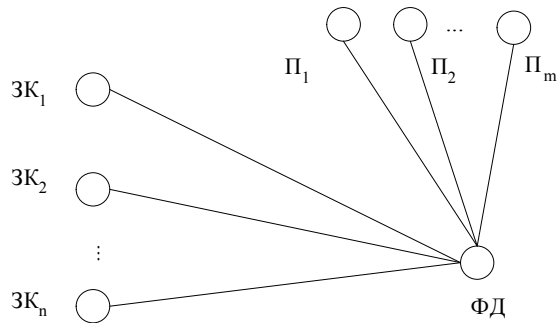


Рис. 4. Фрагмент графа фактографических данных:
 ЗК – значение классификатора; П – показатель;
 ФД – фактографическое данное

Таким образом, вершины фактографических данных, соединяясь с вершинами графов классификаторов и показателей, образуют граф со сложными взаимосвязями (рис. 5).

Для проведения аналитических операций над фактографическими данными с целью вычисления комплексных показателей будут использоваться не все вершины и ребра графа. Например, для расчета суммарной годовой заработной платы в организационной единице «управление корпоративной политики» нужно просуммировать

значения фактографических данных, у которых есть связи с вершинами, представляющими отделы этого управления, и вершиной показателя «заработная плата». Таким образом, из всего графа, представленного на рис. 5, будут использоваться только вершины фактографических данных и связанные с ними вершины графов классификаторов и показателей. Кроме того, если в качестве исходных данных для алгоритма вычисления комплексного показателя будут использоваться множество вершин значений классификаторов и множество вершин показателей, для которых нужно произвести вычисление, то рёбра между вершинами значений классификаторов будут не нужны. Поэтому, учитывая все вышесказанное, упростим граф фактографических данных (см. рис. 5) до трехдольного графа (рис. 6), который представляет собой три множества вершин (X_1, X_2 и X_3), в каждом из которых вершины не имеют рёбер между собой, но имеют рёбра между вершинами соседних множеств [6].

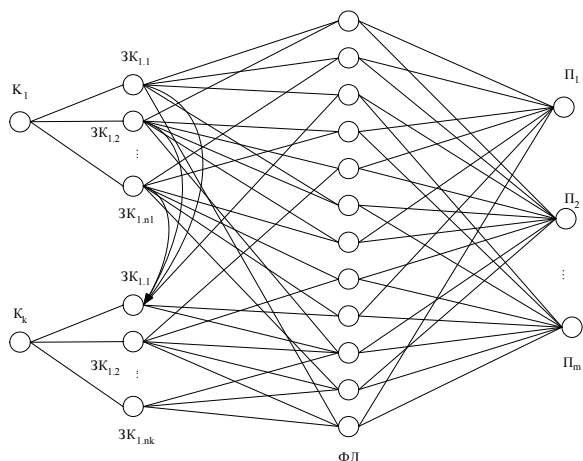


Рис. 5. Граф фактографических данных: К – классификатор; ЗК – значение классификатора; П – показатель; ФД – фактографическое данное

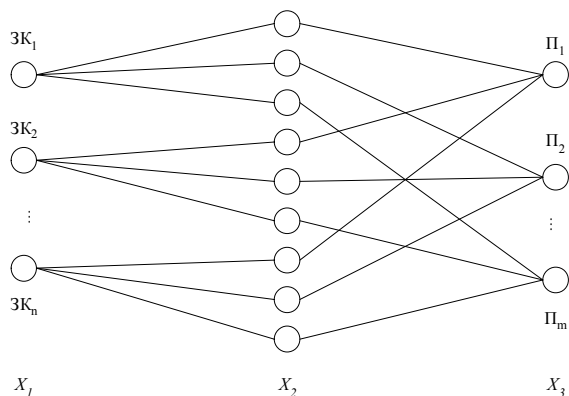


Рис. 6. Трехдольный граф фактографических данных: ЗК – значение классификатора; П – показатель; X – множество вершин трехдольного графа

Трехдольный граф является частным случаем N-дольного графа (рис. 7), который обладает собствен-

ными свойствами отличительными от простого графа, для которого уже существуют некоторые алгоритмы его обхода [7].

Такие алгоритмы плохо работают на N-дольных графах, поскольку они ориентированы на достижение других целей (нахождение кратчайшего пути, поиск определенных значений и др.), а не решения задачи отбора вершин в одной из долей с заданными условиями через подмножества вершин в соседних долях. Именно с такими свойствами требуется алгоритм обхода графа для расчета комплексных показателей. В данной работе предлагаются алгоритмы обхода N-дольного графа с целью поиска вершин в одной доле, имеющих полное множество ребер¹ с заданными через условия поиска подмножествами вершин в соседних долях графа.

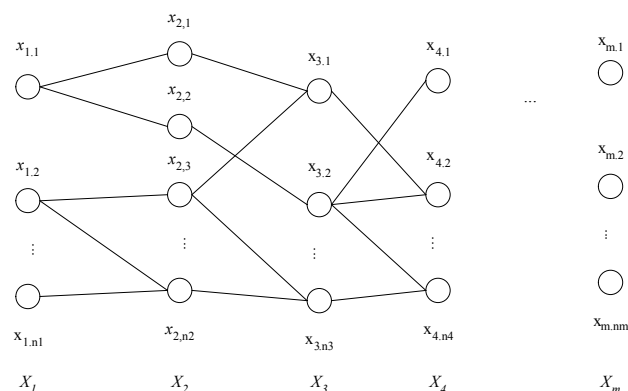


Рис. 7. N-дольный граф:

x – вершина графа;

X – множество вершин трехдольного графа

Алгоритмы поиска на N-дольном графе. Для осуществления поиска вершин в одной доле N-дольного графа необходимо определить два подмножества вершин X'_{k-1} и X'_{k+1} , объединённых в множество условий поиска U_k , где k – номер доли, в которой будет выполняться поиск. При осуществлении поиска вершин в k-й доле будет сформировано подмножество X'_k , в которое будут отобраны вершины, имеющие полное множество ребер с подмножествами из условий поиска:

$$U_k = X'_{k-1} \cup X'_{k+1}. \quad (1)$$

Для каждой вершины $x'_{k,i} \in X'_k$ будет выполняться следующее условие:

$$(\Gamma(x'_{k,i}) \cap X'_{k-1}) \cup (\Gamma(x'_{k,i}) \cap X'_{k+1}) = X'_{k-1} \cup X'_{k+1}, \quad (2)$$

где $\Gamma(x'_{k,i})$ – множество всех вершин, имеющих рёбра с вершиной $x'_{k,i}$.

Ниже описаны два новых алгоритма поиска на N-дольном графе: «обход окрестностей одной вершины» и «пометка вершин».

Обход окрестностей одной вершины

1. Отобрать все вершины из множества X_k , которые имеют рёбра с одним из элементов из множества U_k и поместить их в множество X'_k .

¹ Условиями задаются подмножества вершин для (k – 1) и (k + 1) долей графа, если поиск выполняется в k-й доле. Если вершина k-й доли связана со всеми вершинами заданных подмножеств в (k – 1) и (k + 1) долях, то такая вершина имеет полное множество ребер.

2. Проверить каждую вершину $x'_{k,i} \in X'_k$, имеет ли она полное множество ребер с вершинами множества U_k . Если условие не выполняется, то вершина $x'_{k,i}$ удаляется из множества X'_k .

Анализ данного алгоритма позволил выявить зависимость количества переходов C , необходимых для осуществления поиска, от условий поиска:

$$C = \sum_{i=1}^n (|\Gamma(x'_{k,i})| - 1), \quad (3)$$

где $|\Gamma(x'_{k,i})|$ – количество вершин, связанных с элементом $x'_{k,i}$.

Пометка вершин

1. Выбрать вершину $u_{k,i}$ из множества U_k . Ввести множество меток M , размерность которого равна $|X'_k|$, и каждому элементу этого множества присвоить «0». $i = 1$.

2. Для вершины $u_{k,i}$ найти множество $\Gamma(u_{k,i}) \subset X_k$. Для каждого $x_{k,j} \in X_k$, для которого выполняется условие $x_{k,j} \in \Gamma(u_{k,i})$, соответствующей метке $m_j \in M$ увеличить значение на «1».

3. $i = i + 1$. Если $i = n$, перейти к следующему шагу. Иначе перейти к шагу 2.

4. Найти вершину в множестве U_k с наименьшим количеством ребер и поместить связанные с ней вершины из X_k в множество X'_k .

5. В результирующее множество X'_{res} из множества X'_k отобразить вершины, имеющие полное множество ребер с вершинами из множества U_k : $X'_{res} = \{\forall x_{k,i} \in X'_k \mid m_i = |U_k|\}$.

Определим для данного алгоритма зависимость количества переходов C , необходимых для осуществления поиска, от условий поиска:

$$C = \sum_{i=1}^n |\Gamma(u_{k,i}) \in X_k| + \min(A), \quad (4)$$

где $A = \{|\Gamma(u_{k,i}) \in X_k|, i = \overline{1, n}\}$.

Анализ формул (3), (4) показал, что при небольшом количестве условий поиска (от 1 до 4) эффективнее (с точки зрения количества переходов между вершинами графа) работает алгоритм «пометка вершин», а при увеличении условий поиска эффективнее работает алгоритм «обход окрестностей одной вершины».

Алгоритмы вычисления комплексных показателей.

Комплексный показатель вычисляется на множестве фактографических данных X_2 трехдольного графа G_3 (рис. 6) путем выделения подмножества $X'_2 \subset X_2$, удовлетворяющего условиям U_2 (1), и выполнения операций над этим подмножеством. Таким образом, функция вычисления комплексного показателя (КП) представляет собой алгоритмическую функцию поиска на трехдольном графе с заданными условиями и аналитическую функцию преобразования над полученным множеством:

$$КП = F(G_3, U_2). \quad (5)$$

Рассмотрим пример вычисления комплексного показателя с применением новых алгоритмов поиска на N-дольных графах. Допустим, на трехдольном графе фактографических данных (рис. 8) требуется вычислить количество выпускников в 2007 г. по специальности «токарь». В таблице приведено описание вершин графа для данного примера.

Перед началом поиска в соответствии с примером зададим множество условий (1):

$$U_2 = \{x_{1,3}, x_{1,5}\} \cup \{x_{3,1}\} = \{x_{1,3}, x_{1,4}, x_{3,1}\}.$$

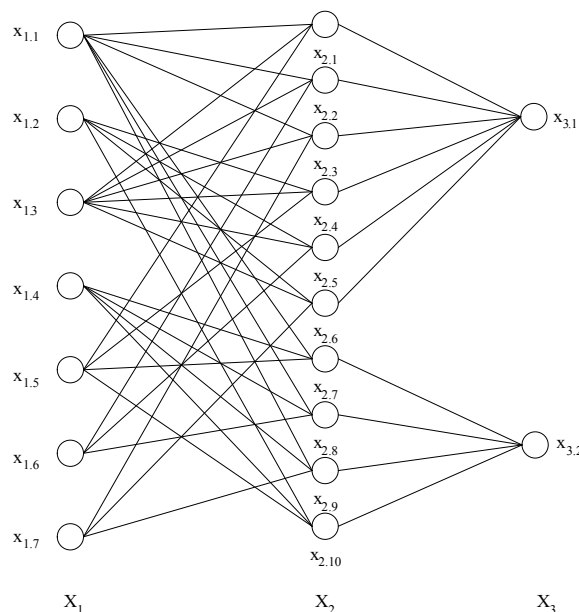


Рис. 8. Пример фактографических данных о выпуске специалистов: x – вершина графа; X – множество вершин трехдольного графа

Применение алгоритма «Обход окрестностей одной вершины»:

1. Отобразить все вершины из множества X_2 , которые имеют ребра с вершиной $x_{1,3}$ из множества U_2 , и поместить их в множество $X'_2 = \{x_{2,1}, x_{2,2}, x_{2,3}, x_{2,4}, x_{2,5}, x_{2,6}\}$.

2. Проверить каждую вершину $x_{2,i} \in X'_2$, имеет ли она полное множество ребер с вершинами множества U_2 . Если условие не выполняется, то вершина $x_{2,i}$ удаляется из множества X'_2 . В результате в множестве X'_2 остаются следующие элементы $\{x_{2,1}, x_{2,3}\}$.

Значения элементов множества X'_2 берем из таблицы. Для вычисления комплексного показателя «количество выпущенных специалистов в 2007 г.» выполним операцию суммирования над элементами получившегося подмножества и получим искомое значение – 15.

Применение алгоритма «Пометка вершин»:

1. $i = 1$. Выбрать вершину $u_{2,i}$ из множества U_2 : ($u_{2,1} = x_{1,3}$). Ввести множество меток M , размерность которого равна $|X_2| = 10$ и каждому элементу этого множества присвоить «0».

2. Для вершины $u_{2,i}$ найти множество $\Gamma(u_{2,i}) \subset X_2$ для $\forall i$. Для каждого $x_{2,j} \in X_2$, для которого выполняется условие $x_{2,j} \in \Gamma(u_{2,i})$, соответствующей метке $m_j \in M$ увеличить значение на «1».

$i = 1$: $u_{2,1} = x_{1,3}$, $\Gamma(u_{2,1}) = \{x_{2,1}, x_{2,2}, x_{2,3}, x_{2,4}, x_{2,5}, x_{2,6}\}$,

$M = \{1, 1, 1, 1, 1, 1, 0, 0, 0, 0\}$;

$i = 2$: $u_{2,2} = x_{1,5}$, $\Gamma(u_{2,2}) = \{x_{2,1}, x_{2,3}, x_{2,10}\}$, $M = \{2, 1, 2, 1, 1, 1, 0, 0, 0, 1\}$;

$i = 3$: $u_{2,3} = x_{3,1}$, $\Gamma(u_{2,3}) = \{x_{2,1}, x_{2,2}, x_{2,3}, x_{2,4}, x_{2,5}, x_{2,6}\}$,

$M = \{3, 2, 3, 2, 2, 2, 0, 0, 0, 1\}$.

3. Найти вершину в множестве U_2 с наименьшим количеством ребер, в нашем случае $x_{1,5}$, и поместить свя-

занные с ней вершины из X_2 в множество $X'_2 = \{x_{2,1}, x_{2,3}, x_{2,10}\}$.

4. В результирующее множество X'_{res} из множества X'_2 отобразить вершины, имеющие полное множество ребер с вершинами из множества U_2 : $X'_{res} = \{\forall x_{2,i} \in X'_2 \mid m_i = |U_2|\} = \{x_{2,1}, x_{2,3}\}$.

Комплексный показатель вычисляется аналогично предыдущему примеру и равняется 15.

Таким образом, предложенный авторами подход представления динамических структур данных позволяет создавать информационные системы с изменяемой структурой информации об объектах предметной области и её обработки на уровне региона. Описанные в статье алгоритмы могут использоваться в разработке автоматизированных систем с динамическими структурами, создаваемых на базе предложенного подхода, для вычисления комплексных показателей при обработке статистической информации.

В развитие данного подхода планируется проведение дальнейших исследований: выявление структуры графа представления показателей, особенностей его построения и алгоритмов обхода этого графа; исследование возможностей хранения графов показателей, классификаторов и фактографических данных в системах управления базами данных в неизменном виде и разработка методик

работы с элементами графов в этих системах; исследование возможных способов автоматической обработки фактографических данных с целью выявления недостоверных данных, новых комплексных показателей и новых классов данных для дальнейшего их анализа.

Библиографические ссылки

1. Постреляционная СУБД Cache 5. Объектно-ориентированная разработка приложений / В. Кирстен, М. Иррингер, М. Кюн, Б. Рериг. 2-е изд. М. : Бином-Пресс, 2005. 416 с.
2. Кайт Т. Oracle для профессионалов. СПб. : ДиаСофт-ЮП, 2005. 656 с.
3. Фаулер М., Скот К. UML. Основы. СПб. : Символ-Плюс, 2002. 192 с.
4. Волкова В. Н., Денисов А. А. Теория систем : учеб. пособие. М. : Высш. шк., 2006. 511 с.
5. Шовкун А. В. Построение корпоративной информационно-аналитической системы в условиях постоянно изменяющегося бизнеса // Науч.-техн. информация. Серия 1. М. : ВИНТИ, 2004. № 9. С. 1–6.
6. Харари Ф. Теория графов. М. : Эдиториал УРСС, 2003. 296 с.
7. Ананий В., Левитин А. Алгоритмы: введение в разработку и анализ = Introduction to The Design and Analysis of Algorithms. М. : Вильямс, 2006. С. 189–195.

Условные обозначения значений классификаторов и показателей

Вершина	Характеристика вершины	Пояснение
$x_{1,1}$	ПУ-1	Наименование образовательного учреждения
$x_{1,2}$	ПУ-2	Наименование образовательного учреждения
$x_{1,3}$	2007	Период (весь 2007 г.)
$x_{1,4}$	2008	Период (весь 2008 г.)
$x_{1,5}$	Токарь	Наименование профессии
$x_{1,6}$	Водитель	Наименование профессии
$x_{1,7}$	Повар	Наименование профессии
$x_{2,1}$	10	Количественная характеристика фактографического данного
$x_{2,2}$	20	Количественная характеристика фактографического данного
$x_{2,3}$	5	Количественная характеристика фактографического данного
$x_{2,4}$	7	Количественная характеристика фактографического данного
$x_{2,5}$	13	Количественная характеристика фактографического данного
$x_{2,6}$	3	Количественная характеристика фактографического данного
$x_{2,7}$	7	Количественная характеристика фактографического данного
$x_{2,8}$	4	Количественная характеристика фактографического данного
$x_{2,9}$	11	Количественная характеристика фактографического данного
$x_{2,10}$	6	Количественная характеристика фактографического данного
$x_{3,1}$	Выпуск специалистов	Наименование показателя
$x_{3,1}$	Набор студентов на первый курс	Наименование показателя

Yu. Yu. Yakunin, A. A. Gorodilov

ALGORITHMS OF CALCULATION OF COMPLEX INDICATORS IN DYNAMIC STRUCTURE OF DATA VIEW

This paper presents algorithms of calculation of complex indicators on a set of factual data display in dynamic structures with application of the graph theory.

Keywords: dynamic structures of the data, tripartite graph, algorithm of graph's round, complex indicators.

© Якунин Ю. Ю., Городилов А. А., 2010