УДК 681.324
Л.П. Луговская, ст. преподаватель,
В.И. Шевченко, ст. преподаватель,
И.А. Скатков, канд. техн. наук, доцент
Севастопольский национальный технический университет
ул. Университетская 33, г. Севастополь, Украина, 99053
E-mail: kvt@sevgtu.sebastopol.ua

ДИНАМИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ИНФОРМАЦИОННЫХ ПОТОКОВ

Рассмотрен подход к решению задачи динамической классификации информационных потоков, циркулирующих в информационных системах. Предложен алгоритм классификации, на основе методов кластерного анализа.

Ключевые слова: *информационные технологии, кластерный анализ, информационная система, информационный поток.*

На этапе эксплуатации информационных систем (ИС) пользователи сталкиваются с рядом проблем, связанных с недостаточным уровнем качества предоставляемых сервисов. Эти проблемы могут быть вызваны изменениями бизнес-процессов предприятия, приводящих к дефициту информационно-вычислительных ресурсов. Инвестиционный подход к модернизации ИС не всегда оправдан. Решение задачи качественной программной настройки (реинжиниринга) ИС позволяет избежать преждевременного аппаратного наращивания системы и, следовательно, избыточных финансовых вложений.

Особенности ИС (функционирование в режиме реального времени, многоуровневая структура, многофазность решаемых задач, динамически меняющаяся интенсивность потока задач, в зависимости от внешних и внутренних факторов воздействия) требуют учета и исследования информационных потоков (ИП), в связи с чем, актуальной является задача разработки моделей управления ИП, циркулирующими в контуре ИС.

Объектом исследования является информационная сеть предприятия, включающая ряд информационных систем неоднородной структуры. Целью исследования является повышение эффективности управления информационным обменом в условиях возможных изменений характеристик ИП и требований к качеству их обработки. В рамках поставленной задачи предлагается подход, позволяющий лицу, принимающему решения, оперативно проводить классификацию потоков данных и на ее основе осуществлять выбор эффективной типовой технологии их обработки.

В работе [1] предложена модель управления обработкой ИП, на основе типовых технологий, согласно которой, при решении задачи по эффективному управлению ИП синтезируется матрица технологий, представленная в виде таблицы 1.

Таблица 1 — Матрица технологий

Технология обработки ИП	Класс ИП			
	I_1	I_2	…	I_K
U_1	F_{11}	F_{12}	…	F_{1K}
U_2	F_{21}	F_{22}	…	F_{2K}
…	…	…	…	…
U_H	F_{H1}	F_{H2}	…	F_{HK}

Каждая строка матрицы соответствует элементу вектора технологий обработки ИП $U = \{u_h \mid h = 1, H\}$. Информационная модель [1] описывается следующими элементами: $\{I\}$ – множество ИП, циркулирующих в ИС, связанных с решением функциональных задач (ФЗ) пользователей; $\{Y\}$ – множество информационных элементов, результат обработки ИП, определяющее решение ФЗ. Качество решения ФЗ оценивается множеством критериев $\{F\}$.

На этапе выбора технологии обработки данных лицо, принимающее решение (ЛПР), использует матрицу технологий для принятия решения о выборе наиболее эффективной технологии обработки данных, с учетом текущих характеристик информационных потоков: $U : I \to Y, F(U, I) \to extr$. Составляющими $\{F\}$ могут быть использованы метрики управления ИТ-услугами [2], такие как: число случаев нарушений SLA, затраты на предоставление услуг (обслуживание ИС), число проблем, возникших при обслуживании функциональных задач (события / инциденты).

Для применения данного подхода ЛПР необходимо провести анализ ИП, с целью выявления классификационных признаков, провести классификацию ИП для выбора типовой технологии обработки наиболее эффективной для данного класса ИП. При проведении классификации необходимо учитывать динамическое изменение характеристик ИП под влиянием внешних воздействий, а так же тот фактор, что ИС, функционирующие в режиме реального времени, требуют оперативного выбора новых технологий обработки при изменении характеристик ИП или возникновении коллизий при обработке ИП.

Для построения и решения задачи классификации ИП необходимо определить ИП и его классификационные признаки.

Определение 1: Адресный информационный поток — совокупность данных, передаваемых между объектами O^I и O^P. В этом случае модель адресного информационного потока имеет вид: $M^1 = \langle O^I, O^p \rangle$, где $\{O^I\}$ — множество объектов (источников информации), $\{O^P\}$ — множество объектов (приемников информации).

Определение 2: Функциональный информационный поток — совокупность данных, передаваемых между объектами O^I и O^P, связанных с решением ФЗ. Модель функционального информационного потока: $M^2 = \langle O^I, O^p, Z \rangle$, где $\{Z\}$ — множество функциональных задач, в процессе решения которых возникает информационный обмен между элементами множества $\{O^I\}$ и $\{O^P\}$.

Определение 3: Динамический информационный поток — информация, передаваемая между объектами O^I и O^P, связанная с решением ФЗ и ограниченная по времени обработки T и объемам передаваемых данных V. Модель динамического информационного потока: $M^3 = \langle O^I, O^p, Z, T, V \rangle$.

На основании приведенных определений можно сделать вывод, что чем выше уровень детализации в описании информационной модели, тем больше классификационных признаков, характеризующих ИП, может быть выделено. На этапе решения задач классификации степень детализации информационной модели ИП определяет ЛПР.

Для проведения классификации в работе использованы методы кластерного анализа, которые широко используется в информационных системах при решении задач классификации и обнаружения закономерностей в данных. Введем необходимые понятия и обозначения, связанные с кластерным анализом и используемые в дальнейших исследованиях.

Согласно [3] кластерный анализ – это способ группировки многомерных объектов, основанный на представлении результатов отдельных наблюдений точками подходящего геометрического пространства с последующим выделением групп, как сгущений этих точек. Основная цель анализа — выделить в исходных многомерных данных такие однородные подмножества, чтобы объекты внутри групп были похожи в известном близки, согласно выбранной метрики близости. Пусть имеется множество из n объектов исследования $\{M_n\}$, каждый из которых описывается некоторым множеством наблюдаемых и измеряемых показателей или характеристик. Требуется сформировать $K \geq 2$ классов (групп объектов). Число классов может быть, как выбрано заранее, так и не задано (в последнем случае оптимальное количество кластеров должно быть определено автоматически). Результаты измерения *j*-й характеристики *i*-го объекта обозначаются x_{ij}, а вектор размерности R в этом случае будет отвечать каждому ряду измерения (для *i*-го объекта). Каждый объект в этом случае можно интерпретировать при этом как точку R–мерного пространства с координатами, равными значениям признаков для рассматриваемого объекта. По результатам наблюдений формируется матрица $X[n, R]$, где n — число объектов, R — число показателей.

Задача кластерного анализа ИП заключается в том, чтобы на основании данных, содержащихся в матрице *X*, разбить множество $\{M_n\}$ на подмножества так, чтобы каждый ИП принадлежал одному и только одному подмножеству разбиения (непересекающиеся подмножества). При этом, чтобы ИП, принадлежащие одному и тому же подмножеству, были сходными, в то время как ИП, принадлежащие разным подмножествам, были разнородными (несходными). Можно выделить следующие основные этапы кластерного анализа:

1. *Формирование системы классификационных признаков.* Часто ЛПР не может с уверенностью сказать, какие именно классификационные действительно важны для анализа, поэтому стремится включить как можно больше потенциально информативных факторов. Нередко требуется предварительно выбрать из исходного множества признаков наиболее эффективные («feature selection»). Кроме того, в некоторых задачах целесообразно трансформировать исходные признаки так, чтобы образовать новые, более информативные показатели («feature extraction»). Чтобы избежать

«доминирования» классификационных признаков с большим масштабом измерения, проводят предварительную нормировку исходных признаков.

Наиболее распространенными способами стандартизации показателей являются [3]:

$$x_{ij}^{cm} = \frac{x_{ij}}{\overline{x}_j} \;;\; x_{ij}^{cm} = \frac{x_{ij}}{x_j^{эт}} \;;\; x_{ij}^{cm} = \frac{x_{ij}}{x_j^{\max}} \;;\; x_{ij}^{cm} = \frac{x_{ij} - \overline{x}_j}{x_i^{\max} - x_i^{\min}},$$

где $\overline{x}_j$ – среднеарифметическое значение j-го признака; $x_j^{эт}$ — некоторое эталонное значение j-го признака; $x_j^{\max}$ – максимальное значение j-го признака по всем i-м объектам; $x_{ij} - \overline{x}_j$ – центрированное значение признака; $x_j^{\min}$ – минимальное значение j-го признака по всем i-м объектам;

Достаточно часто используется схема, где соответствующее преобразование производится в соответствии с формулой:

$$x_{ij}^{cm} = \frac{x_{ij} - \overline{x}_j}{\sigma_{x_j}},$$

где σ_{x_j} — среднеквадратическое отклонение по всем объектам.

В качестве классификационных признаков ИП могут быть использованы: адресные признаки — адреса объекта (процедуры) источника данных и приемника данных, функциональные признаки по типам решаемых задач, отношению к бизнес-процессам, интенсивность поступления данных, средний объем передаваемой информации, среднее время обработки информации, тип обработки (чтение, запись, создание, корректировка), тип данных (электронный текстовый документ, бумажный документ, видео информация, аудио информация и т. п.).

2. *Определение способа вычисления расстояния между объектами или группами объектов.* Этот способ должен отражать специфику решаемой прикладной задачи. Для каждой пары объектов x_{ij} и x_{kj} обозначим расстояние между ними $d_{ik}, i \neq k$. В качестве метрики расстояния может быть использовано Эвклидово расстояние, которое рассчитывается по формуле: $d_{ik}^2 = \sum_{j=1}^{R} \left(x_{ij} - x_{kj}\right)^2$.

3. *Группировка объектов.* На этом шаге создаются группы информационных потоков. Разбиение на группы может быть «жестким» (формируется разбиение исходного множества объектов), а может быть нечетким (вычисляется степень принадлежности каждого объекта к группам). В данной работе будем рассматривать группировку первого типа.

4. *Определение качества полученной группировки*. Лицу, принимающему решения, необходимо удостовериться в том, что сформированные группы действительно отражают внутренние закономерности, характерные для решаемой задачи. Существуют формальные способы проверки качества, связанные с нахождением вероятности случайного образования групп, которую можно вычислить в рамках той или иной модели распределения [4]. В качестве критериев оценки классификации могут быть использованы [3]:

– общая внутриклассовая дисперсия по всем признакам: $F_1 = \sum_{i=1}^{nc} \sum_{j=1}^{R} \sigma_{ij}^2$;

– отклонения от центров классов: $F_2 = \sum_{j=1}^{R} \sum_{x_i \in C} (x_i - \overline{x})$;

– отношение средних внутриклассовых и межклассовых расстояний: $F_3 = \left(d_W / f_W\right) / \left(d_b / f_b\right)$.

Суть предлагаемого подхода заключается в следующем. Существует ряд алгоритмов, позволяющих производить кластеризацию объектов по заданным признакам [3, 4], эти алгоритмы не учитывают специфики обработки ИП в ИС. Специфика связана с ограничением времени обработки, так как большинство задач решаются в режиме реального времени. В связи с этим, иерархические алгоритмы кластерного анализа не могут быть применены при решении задачи выбора альтернативной технологии обработки ИП. В рамках предлагаемого подхода, на первом этапе лицо, принимающее решения, задает, основываясь на экспертных оценках информацию о составе классификационных признаков информационных потоков, по которым производится кластеризация, количестве кластеров nc, их центрах. Затем, с помощью методов многомерной оптимизации, определяется эффективное расположение центров кластеров относительно выбранной метрики. На основе полученных данных ЛПР выбирает наиболее эффективную технологию обработки ИП из матрицы технологий. Общая схема

коррекции центров кластеров и алгоритм коррекции центра j-го кластера по критерию минимума среднего расстояния до центра приведены на рисунках 1 и 2.

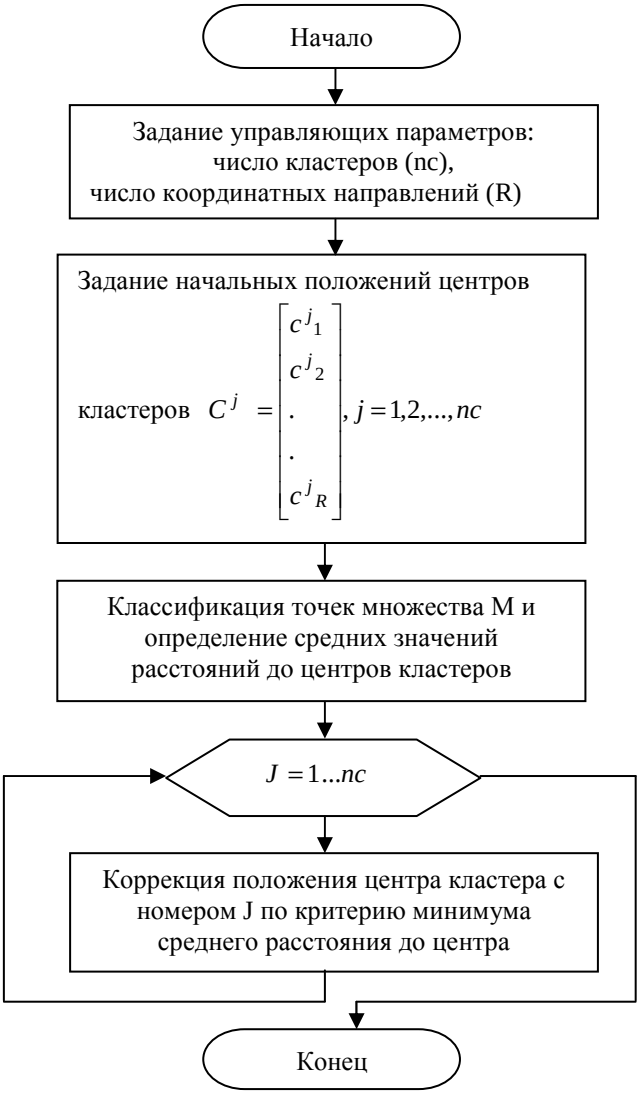

Рисунок 1 — Общая схема алгоритма коррекции положения центров кластеров

Рассмотрим работу алгоритма на примере. Предполагаем, что ИП оцениваются по классификационным признакам x_1 и x_2. Матрица технологий содержит информацию о трех классах ИП, обрабатываемых в ИС (потоки нормативно-справочной информации, потоки данных аналитической отчетности и потоки, связанные с текущим решением задач в диалоговом режиме). Лицом, принимающим решения, сделано предположение о начальном положении центров кластеров для каждого вида ИП. Согласно схеме (рисунок 2) производится коррекция первоначального положения центров кластеров по алгоритму Хука-Дживса. В качестве метрики эффективности использован минимум среднего расстояния до центра кластера. Данные о центрах кластеров до и после процедуры коррекции приведены в таблице 1 и показаны на рисунке 3.

Таблица 1 — Данные о центрах кластеров до и после коррекции

Положение кластера	№ кластера	Координаты центров кластера		Число точек, отнесенных к кластеру	Среднее расстояние до центра
Начальное положение	1	9,0	2,0	18	4,7
	2	11,0	4,0	38	10,0
	3	13,0	4,0	44	5,9
После коррекции	1	5,0	4,0	20	2,4
	2	7,0	13,0	39	3,6
	3	18,0	3,0	41	2,4

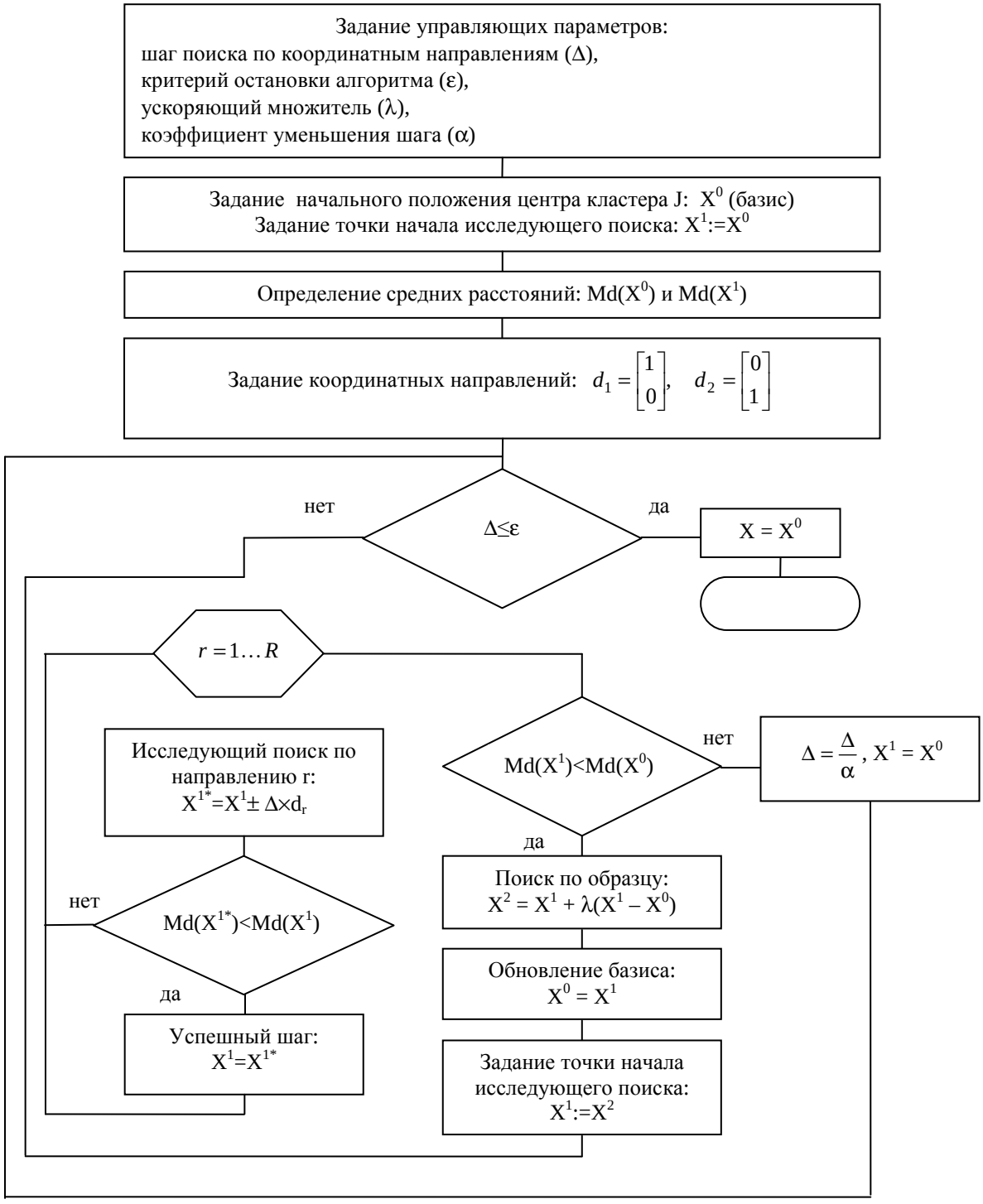

Рисунок 2 – Коррекция положения центра кластера по критерию минимума среднего расстояния до центра

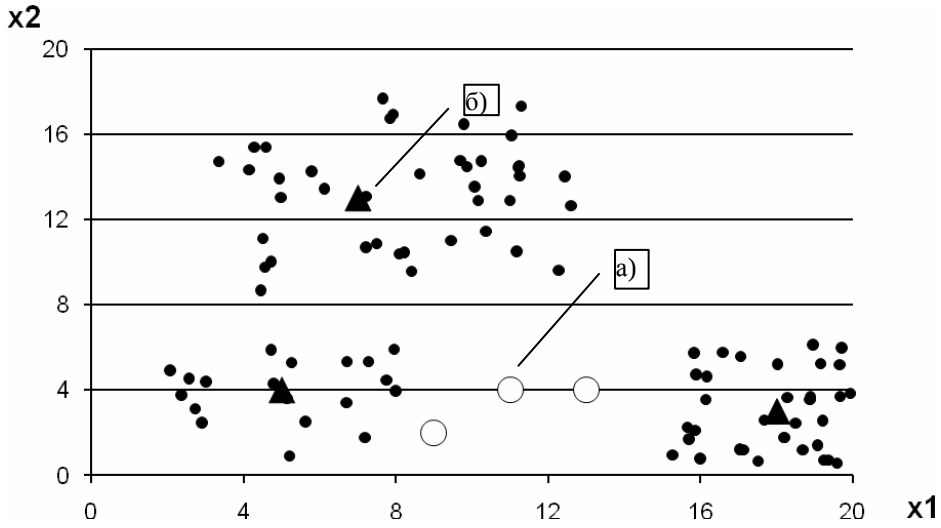

Рисунок 3 — Данные о центрах кластеров до и после коррекции:
а) начальное положение центров кластеров; б) положение центров после коррекции

В качестве критерия оценки классификации в работе использован минимум среднего расстояния до центров кластеров. На рисунке 4 показана зависимость среднего расстояния до центра кластеров от числа итераций алгоритма.

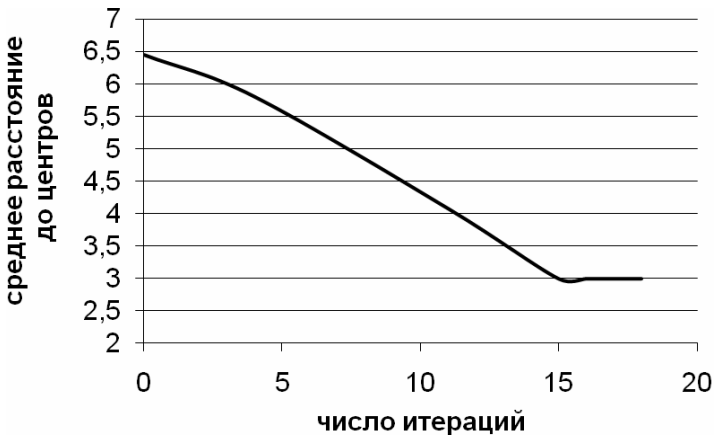

Рисунок 4 — Зависимость среднего расстояния до центров кластеров от числа итераций

Анализ полученных результатов позволяет сделать следующие выводы:

1) предлагаемый подход можно использовать для принятия решения по выбору эффективного с точки зрения лица принимающего решение варианта управления информационными потоками в КИС по заданным критериям эффективности;
2) предлагаемый алгоритм коррекции центров кластеров эффективен при решении задач оптимизации функционирования реальных корпоративных информационных сетей.
3) для использования подхода необходимо собрать достаточный объем статистических данных об обрабатываемых в КИС информационных потоках.

В дальнейших исследованиях предполагается расширить комплекс моделей управления распределением информационных потоков для различных стратегий управления и структур КИС в условиях динамически меняющихся характеристик информационных потоков. На основе полученного комплекса моделей предполагается создание системы поддержки принятия решений, по эффективному управлению распределением и обработкой информационных потоков в КИС.

Библиографический список использованной литературы

1. Шевченко В.И. Многоуровневая графовая модель информационного обмена в корпоративной информационной сети / В.И. Шевченко // Оптимизация производственных процессов: сб. науч. тр. — Севастополь, 2009. — Вып. № 11. — С. 201–205.
2. Брукс П. Метрики для управления ИТ-услугами / Питер Брукс; пер. с англ. — М: Изд-во «Альпина Бизнес Букс», 2008. — 283 с.
3. Мандель И.Д. Кластерный анализ / И.Д. Мандель. — М.: Финансы и статистика, 1988. — 176 с.
4. Дюран Б. Кластерный анализ / Б. Дюран, П. Одел. — М.: Статистика, 1977. — 128 с.

Поступила в редакцию 01.10.2010 г.

Луговская Л.П., Шевченко В. І., Скатков І.А. Динамічна кластеризація інформаційних потоків

Розглянутий підхід до рішення завдання динамічної класифікації інформаційних потоків, циркулюючих в інформаційних системах. Запропонований алгоритм класифікації на основі методів кластерного аналізу.

Ключові слова: інформаційні технології, кластерний аналіз, інформаційна система, інформаційний потік.

Lugovskaya L.P., Shevchenko V.I., Scatkov I.A. Dynamic cauterization of information flows

An approach to the solution of dynamic classification task for the information flows circulating in information systems is considered. On the basis of cluster analysis, the algorithm of classification is offered.

Keywords: Information technologies, cluster analysis, information system, information flow.