7. Лившиц, Б. Г. Физические свойства металлов и сплавов [Текст] / Б. Г. Лившиц, В. С. Крапошин, Я. Л. Линецкий. – М : "Металлургия", 1980. – 320 с.

8. Князева, А. Г. Проблемы моделирования технологических процессов поверхностной обработки материалов и нанесения покрытий с использованием высокоэнергетических источников [Текст] / А. Г. Князева, О. Н. Крюкова, Н. В. Букрина, С. Н. Сорокова // Известия ТПУ. – 2010. – № 2. – С. 93–101.

9. Иванов, Д. А. Расчет теплоемкости низкоуглеродистой низколегированной стали при моделировании неизотермических фазовых превращений [Текст] / Д. А. Иванов, Н. В. Куваев, Т. В. Куваева // Теория и практика металлургии. – 2010. – № 1–2. – С. 43–48.

10. Комп'ютерне моделювання у лазерних технологіях [Текст] / Л. Ф. Головко, С. О. Лук'яненко, І. Ю. Михайлова, В. А. Третяк. – К. : ВПП "Текст", 2015. – 236 с.

11. Бадаєв, Ю. І. Реалізація інтерполяційного методу Гаусс-функції та порівняльний аналіз [Текст] / Ю. І. Бадаєв, Ю. В. Сидоренко. Прикладна геометрія та інженерна графіка. – 1998. – Вип. 63. – С.33–37.

**References**

1. Verhoeven, J. C. J., Jansen, J. K. M., Mattheij, R. M. M., Smith, W. R. (2003). Modelling laser induced melting. Mathematical and Computer Modelling, 37 (3-4), 419–437. doi: 10.1016/s0895-7177(03)00017-7

2. Solov'eva, E. N., Uspenskiy, A. B. (1975). Skhemy skvoznogo scheta chislennogo resheniya kraevyh zadach s neizvestnymi granitsami dlya odnomernyh uravnenij parabolicheskogo tipa [Schemes of through computation of the numerical solution of boundary value problems with unknown boundaries for one-dimensional parabolic equations]. Methods of solving boundary and inverse heat conduction problems, 5, 3–23.

3. Breslavskiy, P. V., Mazhukin, V. I. (1991). Algoritm chslennogo resheniya gidrodinamicheskogo varianta zadachi Stefana pri pomoshchi dinamicheski adaptiruyushchihsya setok [The algorithm of a hydrodynamical version of Stefan problem numerical solution by dynamic adapting grid]. Mathematical modeling, 3 (10), 104–115.

4. Luk'yanenko, S. A., Tretyak, V. A. (2014). Problema ucheta zavisimosti koeffitsienta ob'emnoy teploemkosti ot temperatury pri modelirovanii lazerno-dugovoy naplavki [Temperature dependence consideration issue for coefficient of volumetric heat capacity in simulation of laser-arc pad weld process]. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 1 (89), 177–182.

5. Pereloma, V. A., Shcherba, A. A., Podol'tsev, A. D. et al. (1998). Issledovanie teplovyh protsessov i structury poverhnostnogo sloya pri lazernoy naplavke poroshkovyh materialov [Heat process and the cover structure while laser cladding of powder materials research]. The Institute of Electrodynamics of the National Academy of Sciences of Ukraine, 47.

6. Amara, E. H., Hamadi, F., Achab, L., Boumia, O. (2006). Numerical modelling of the laser cladding process using a dynamic mesh approach. Journal of Achievements in Materials and Manufacturing Engineering, 15 (1-2), 100–106. doi: 10.1109/caol.2005.1553842

7. Livshits, B. G., Kraposhin, V. S., Linetskiy, Ya. L. (1980). Fizicheskie svoystva metallov i splavov [Physical properties of metals and alloys]. Metallurgy, 320.

8. Knyazeva, A. G., Kryukova, O. N., Burkina, N. V., Sorokova, S. N. (2010). Problemy modelirovaniya tehnologicheskikh protsessov poverkhnostnoy obrabotki materialov i naneseniya pokrytiy s ispol'zovaniem vysokoenergeticheskikh istochnikov [Simulation issues of surface treatment and coating materials using high energy sources]. Izvestiya of TPU, 317 (2), 93–101.

9. Ivanov, D. A., Kuvaev, N. V., Kuvaeva, T. V. (2010). Raschet teploemkosti nizkouglerodistoy nizkolegirovannoy stali pri modelirovanii neizotermicheskikh fazovikh prevrashcheniy [The heat capacity of low-carbon low-alloy steel calculation for modeling of non-isothermal phase transitions]. Theory and practice of metallurgy, 1–2, 43–48.

10. Holovko, L. F., Lukianenko, S. O., Mikhailova, I. Yu., Tretiak, V. A. (2015). Kompyuterne modeliuvannia u lazernykh tekhnolohiiakh [Computer simulation in laser technologies]. Text, 236.

11. Badaiev, Yu. I., Sydorenko, Yu.V. (1998). Realizatsiia interpoliatsiynoho metodu Gaus-funktsii ta porivnialnyy analiz [Interpolation method by Gauss function implementation and comparative analysis]. Applied geometry and engineering graphics, 63, 33–37.

*Рекомендовано до публікації д-р техн. наук, професор Аушева Н. М.*
*Дата надходження рукопису 23.04.2015*

**Сидоренко Юлія Всеволодівна**, кандидат технічних наук, доцент. кафедра автоматизації проектування енергетичних процесів та систем, Національний технічний університет України «Київський політехнічний інститут», пр. Перемоги, 37, м. Київ, Україна, 03056
E-mail: suliko3@ukr.net
**Третьяк Валерія Анатоліївна**, кандидат технічних наук, доцент, кафедра автоматизації проектування енергетичних процесів та систем, Національний технічний університет України «Київський політехнічний інститут», пр. Перемоги, 37, м. Київ, Україна, 03056
E-mail: valery.tretyak@gmail.com

**УДК 665.9**
**DOI: 10.15587/2313-8416.2015.42641**

## ПОСТРОЕНИЕ ГРАФА СВЯЗНОСТИ В АЛГОРИТМЕ КЛАСТЕРИЗАЦИИ СЛОЖНЫХ ОБЪЕКТОВ

**© Т. Б. Шатовская, И. В. Каменева**

*В статье рассмотрены результаты работы модификации алгоритма Хамелеон. Иерархический многоуровневый алгоритм состоит из нескольких фаз: построение графу, огрубление, разделение и восстановления. На каждой фазе могут быть использованы различные подходы и алгоритмы. Главной целью работы является исследование качества кластеризации различных наборов данных с помощью набора комбинаций алгоритмов на разных этапах работы алгоритма и улучшения этапа построения через оптимизации алгоритма выбора k при построении графа k ближайших соседей*

***Ключевые слова:*** *кластеризация, алгоритм Хамелеон, построение графа, связность, k-ближайших соседей, иерархическая кластеризация*

*The article describes the results of modifying the algorithm Chameleon. Hierarchical multi-level algorithm consists of several phases: the construction of the count, coarsening, the separation and recovery. Each phase can be used various approaches and algorithms. The main aim of the work is to study the quality of the clustering of different sets of data using a set of algorithms combinations at different stages of the algorithm and improve the stage of construction by the optimization algorithm of k choice in the graph construction of k of nearest neighbors*

***Keywords:*** *clustering, algorithm Chameleon, graph construction, connectivity, k-nearest neighbors, hierarchical clustering*

**1. Введение**

На данный момент весьма активно исследуются различные методы кластеризации. Каждым из целого множества имеющихся методов можно получить различные разбиения исходного множества. Выбор определенного метода зависит от типа желаемого результата. Производительность метода с определенными типами данных зависит от характеристик сервера и технических возможностей программного обеспечения, размера множества. Модификация алгоритма построения графа в алгоритме Хамелеон.

Целью работы является улучшение этапа построения графа посредством оптимизации алгоритма выбора k при построении графа k ближайших соседей.

Главными задачами являются исследование и совершенствование этапа построения графа путем модификации алгоритма Хамелеон.

**2. Литературный обзор**

Практическая ценность исследования заключается в том, что созданная модель выбора k для алгоритмов k-ближайших соседей, основанная на характеристиках данных для дальнейшего использования в рамках алгоритма Хамелеон [1]. Для ускорения работы алгоритма будет модифицирован алгоритм построения графа, для чего построена математическая модель зависимости k от статических характеристик выборки

В последнее время ведутся активные разработки новых алгоритмов кластеризации, способных обрабатывать сверхбольшие базы данных. В них основное внимание уделяется масштабируемости. Разработаны алгоритмы, в которых методы иерархической кластеризации интегрированы с другими методами. К наиболее актуальным алгоритмам относятся: BIRCH, CURE, CHAMELEON, ROCK [2].

**3. Модифицированный алгоритм Хамелеон**

Хамелеон – это новый иерархический алгоритм, который преодолевает ограничения существующих алгоритмов кластеризации. Данный алгоритм рассматривает динамическое моделирование в иерархической кластеризации. В нем можно выделить следующие стадии:

1. Построение графа. Граф может быть построен симметричный или ассиметричный. Различные виды расстояний могут быть применены при построении графа: Euclidian, Manhattan, Minkowski, SquEuclidian.

2. Огрубление графа (Coarsening). Огрубление графа может быть выполнено следующими методами: Random Matching(RM), Heavy Edge Matching(HEM), Light Edge Matching(LEM).

3. Начальное разделение графа (Initial Partitioning). Существует несколько подходов к разделению графов: графические методы, комбинаторные методы и спектральные методы. Также алгоритмы могут быть выполнены в рамках рекурсивной бисекции, так как большинство методов выполняет деление графа пополам.

4. Восстановление графа (Uncoarsening) и усовершенствование разделения графа (Refinement). Для улучшения разделения графа применяются следующие алгоритмы: Kernighan–Lin (KL), Boundary KL, Fiduccia-Mattheyses (FM), BoundaryFM. Эти же алгоритмы могут быть применены на этапе разделения, взяв за начальное случайное разделение огрубленного графа.

5. Объединение схожих классов для получения финального разбиения.

Целью построения графа является соединение точек локальных соседей. Точки соединяются в зависимости от типа графа.

- Граф эпсилен-окрестности (Epsilon-neighborhood graph). Две вершины графа соединены, если расстояние между рассматриваемыми объектами меньше эпсилен. Данный граф может быть взвешенным и не взвешенным. В случае взвешенного графа вес ребра равняется значению схожести соседних точек (не расстоянию). Параметр данного графа – эпсилен – устанавливается пользователем.
- Полностью связный граф (completely connected graph). Граф может быть получен из графа эпсилен-окрестности установкой эпсилен в максимальное значение.
- Симметричный граф k ближайших соседей (symmetric k-nearest neighbor graph(k-nn)): две вершины x, y соединены, если x находится среди k ближайших соседей y и наоборот.
- Ассиметричный граф k ближайших соседей (mutual k-nearest neighbor graph): две вершины x, y соединены, если x находится среди k ближайших соседей y или наоборот [2].

**4. Введение в k-nn граф**

Задача графа k ближайших соседей определена следующим образом: дано множество точек P из n точек в $R^d$ и положительное целое число $k \le n-1$, рассчитать k ближайших соседей для каждой точки P. Более формально задача может быть представлена следующим образом: пусть $P=\{p_1,p_2,\dots p_n\}$ множество точек в пространстве $R^d$ где $d \le 3$. Для каждой вершины $p_i \in P$ пусть $N^k_i$ k точек из P ближайших к $p_i$. Граф k ближайших соседей (k-nearest neighbor graph

(k-NNG)) – это граф где множество вершин $\{p_1,p_2,\ldots p_n\}$ и множество ребер E= $\{(p_i, p_j)$: $p_i \in N^k_i$ или $p_j \in N^k_i\}$ [3]. Следует отметить, что это ассиметричный граф k ближайших соседей, так как отношения близости ассиметричны. $P_i$ может быть среди ближайших соседей $p_j$, но $p_j$ нет. В симметричном графе $p_i$ и $p_j$ будут соединены ребром только в том случае, если каждая из них находится среди k ближайших соседей другой вершины.

В данной работе рассмотрено 2 вида графов: симметричный k-nn граф и ассиметричный k-nn граф.

При построении графа для каждой пары объектов измеряется «расстояние» между ними – степень похожести.

В данном случае, чем больше сходство меду двумя объектами – тем тяжелее будет ребро между ними.

Еще одним важным параметром при построении графа является k – количество соседей, с которыми будет связанна каждая из вершин. Граф называется *связным*, если в нем для любых двух вершин имеется маршрут, соединяющий эти вершины. При решении поставленной задачи для построения графа k должно быть выбрано таким образом, чтобы соблюдалось условие связности построенного графа. Но слишком большое значение k очень сильно увеличивает вычислительную дороговизну метода и время выполнения не только этапа построения графа, а и всех последующих этапов. Самым простым подходом для выбора k является (1), но и данный метод имеет вышеперечисленные недостатки.

$$k = \sqrt{n}. \quad (1)$$

На практике применяется два принципиально различных порядка обхода, основанных на поиске в глубину и поиске в ширину соответственно.

Поиск в ширину. Вначале все вершины помечаются как новые. Первой посещается вершина a, она становится единственной открытой вершиной. В дальнейшем каждый очередной шаг начинается с выбора некоторой открытой вершины x. Эта вершина становится активной. Далее исследуются ребра, инцидентные активной вершине. Если такое ребро соединяет вершину x с новой вершиной y, то вершина y посещается и превращается в открытую. Когда все ребра, инцидентные активной вершине, исследованы, она перестает быть активной, и становится закрытой. Если на данном этапе остались незакрытые вершины –то граф несвязный.

Поиск в глубину. Главное отличие от поиска в ширину состоит в том, что при поиске в глубину в качестве активной выбирается та из открытых вершин, которая была посещена последней. Основной алгоритм тот же, что и в случае поиска в ширину, только нужно очередь заменить стеком, а процедуру BFS – процедурой DFS.

Общая оценка трудоемкости для алгоритмов одинаковая – $O(m+n)$.

**5. Оптимизация выбора k для построения k-nn графа**

Для оптимизации выбора начального параметра k при построении k-nn графа необходимо построить математическую модель зависимости k от характеристик обрабатываемой выборки. Построение математической модели выполнялось на наборе экспериментальных выборок. Набор выборок состоит из 132 выборок, среди них 33 уникальных выборки и 3 вариаций каждой из них полученной путем добавления 20 %, 40 % и 60 % шума. Эксперимент так же проводился на наборах экспериментальных и реальных выборок полученных с ресурсов обмена наборами данных.

Целью данных экспериментов был выбор управляемых параметров данной модели зависимости, способных отобразить необходимые характеристики выборки данных. В рамках работы было проведено 3 эксперимента для выбора управляемых параметров.

– В первом эксперименте анализировались такие характеристики как: количество объектов в выборке, минимальные и максимальные значения математического ожидания, дисперсии и разброса. Зависимости между данными параметрами и значением k не выявлено.

– Во втором эксперименте в качестве управляемого параметра были выбраны длина наибольшего остового ребра полносвязного графа и среднее значение длины всех остальных ребер остова. Данные характеристики показывают зависимость, но использование данного подхода не является целесообразным в связи с трудоемкостью построения остова полносвязного графа.

– В третьем эксперименте в качестве характеристики использовались количество компонентов связности, максимальное расстояние между компонентами связности и количество элементов в компоненте связности.

В результате исследования была построена математическая модель для оптимизации выбора начального значения k при построении ассиметричного k-nn графа. Модель для ассиметричного k-nn графа имеет следующий вид (2) и представлена на (рис. 1):

$$k = a + b\cdot x_1 + c\cdot x_2 + d\cdot x_1^2 + e\cdot x_2^2 + f\cdot x_1\cdot x_2 + \\ + g\cdot x_1^3 + h\cdot x_2^3 + i\cdot x_1\cdot x_2^2 + j\cdot x_1^2\cdot x_2, \quad (2)$$

где $x_1$ – коэффициент расстояния; $x_2$ – количество компонент связности.

О качестве построенной модели можно судить, исходя из следующих характеристик: стандартная ошибка оценки равна 11,2986020522291, коэффициент множественной детерминации равен 0,6452864929, статистика Дублина-Ватсона составляет 1,24157318003058.

Оценки и статистики качества данной модели не являются остаточными показателями эффек тивности применения полученной модели, так как модель является лишь одним из этапов выбора k.

Применение подхода исследовалось на 285 выборках. Применение данной модели улучшили время выполнения этапа построения графа в 62,45 % случаев. В 37,55 % случаев время выполнения ухудшилось.

Время выполнения ухудшилось лишь в тех случаях, когда k было меньше или равно 3 и время выполнения мало, следовательно, ухудшение временного показателя несущественно сказывается на производительности метода в целом. Отрицательный результат применения модели получен в 7,71 % случаев. В среднем время выполнения улучшилось на 161 %. Отрицательным результат считается при получении k существенно большем минимально необходимого для соблюдения условия связности, даже если время построения графа уменьшилось. Так же в результате исследования была построена математическая модель для оптимизации выбора начального значения k при построении симметричного k-nn графа. Модель для ассиметричного k-nn графа имеет следующий вид (3) и представлена на (рис. 2):

$$k = a + b \cdot x_1 + c \cdot x_1^2 + d \cdot x_1^3 + \\ + e \cdot x_2 + f \cdot x_2^2 + g \cdot x_2^3 + h \cdot x_2^4 + i \cdot x_2^5, \quad (3)$$

где $x_1$ – коэффициент расстояния; $x_2$ – количество компонентов связности.

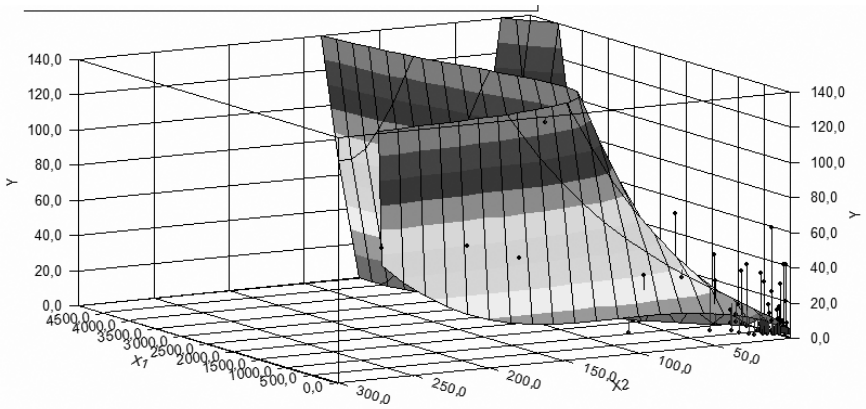


Рис. 1. Графическое представление описания данных математической моделью

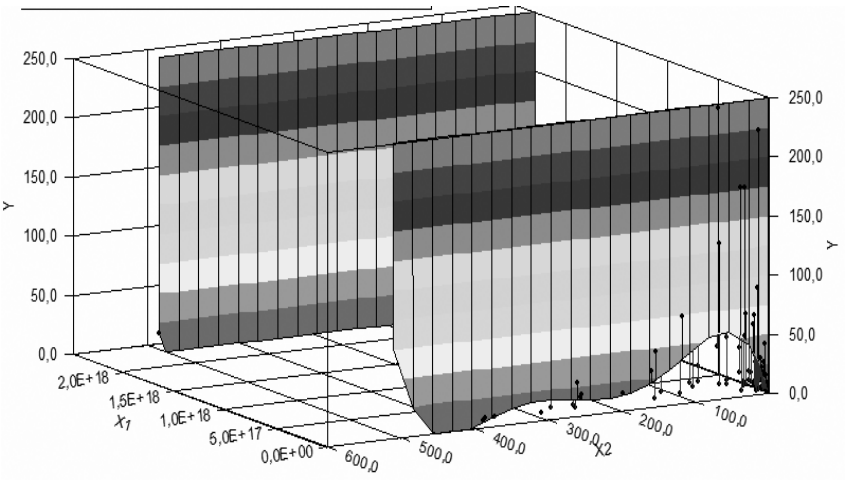


Рис. 2. Графическое представление описания данных математической моделью

О качестве построенной модели можно судить, исходя из следующих характеристик: стандартная ошибка оценки равна 42,8805641130193, коэффициент множественной детерминации равен 0,15118817, статистика Дублина-Ватсона составляет 1,26055939255469.

Применение данной модели улучшило время выполнения этапа построения графа в 69,23 % случаев. В 20,51 % случаев время выполнения ухудшилось. Отрицательный результат применения модели получен в 5,12 % случаев. В среднем время выполнения улучшилось на 169 %.

Использование модели особенно критично для больших выборок. Полученные результаты будут использованы для дальнейших исследований и модификаций алгоритма Хамелеон.

**6. Выводы**

В данной статье разработана математическая модель для выбора алгоритмов в рамках модифицированного алгоритма Хамелеон, построена математическая модель для выбора k при построении k-nn графа в рамках модифицированного алгоритма Хамелеон, приведены результаты применения разработанных методов на реальных данных.

Была разработана математическая модель для выбора k при построении k-nn графа в рамках модифицированного алгоритма Хамелеон. Приведены результаты, полученные как на экспериментальных, так и на реальных данных.

Результаты данной работы позволили усовершенствовать этап построения графа путем модификации алгоритма Хамелеон с целью улучшения процессов кластеризации, ориентированных на работу с очень большими базами данных.

**Литература**

1. Asuncion, A. UCI Machine Learning Repository [Electronic resource] / A. Asuncion, D. J. Newman. – University of California, School of Information and Computer Science, Irvine, CA, 2007. – Available at: http://www.ics.uci.edu/~mlearn/MLRepository.html

2. Blake, C. L. UCI repository of machine learning databases [Electronic resource] / C. L. Blake, C. J. Mer. – 2001. – Available at: http://www.ics.uci.edu/~mlearn/ML-Repository.html

3. Pearson, S. An Adaptive Privacy Management System For Data Repositories [Electronic resource] / S. Pearson, M. Mont, P. Bramhall. – Trusted Systems Laboratory, Hewlett-Packard Laboratories, Bristol, UK, 2004. – Available at: http://www.hpl.hp.com/techreports/2004/HPL-2004-211.pdf

4. Cunningham, K. An open repository and analysis tools for fine-

grained longitudinal learner data [Electronic resource] / K. Cunningham, R. Kenneth, Koedinger, A. Skogsholm, B. Le-ber. – Human Computer Interaction Institute, Carnegie Mellon University, 2008. – Available at: http://www.educationaldatamining.org/EDM2008/uploads/proc/16_Koedinger_45.pdf

5. Xie T. JMAPO: mining API usages from open source repositories. [Electronic resource] / T. Xie, J. Pei // Proceedings of the International Workshop on Mining Software Repositories (MSR '06)ACM. – Press, New York. Shanghai, Chinapp, 2006. – P. 54–57. Available at: http://people.engr.ncsu.edu/txie/publications/msr06-mapo.pdf doi: 10.1145/1137983.1137997

6. Zimmermann, T. Knowledge Collaboration by Mining Software Repositories [Electronic resource] / T. Zimmermann. – Saarland University, Saarbrucken, Germany, 2006. – Available at: http://thomas-zimmermann.com/publications/files/zimmermann-kcsd-2006.pdf

**References**

1. Asuncion, A., Newman, D. J. (2007). UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA. Available at: http://www.ics.uci.edu/~mlearn/MLRepository.html

2. Blake, C. L., Mer, C. J. (2001). UCI repository of machine learning databases. Available at: http://www.ics.uci.edu/~mlearn/ML-Repository.html

3. Pearson, S., Mont, M., Bramhall, P. (2004). An Adaptive Privacy Management System For Data Repositories. Trusted Systems Laboratory, Hewlett-Packard Laboratories, Bristol, UK. Available at: http://www.hpl.hp.com/techreports/2004/HPL-2004-211.pdf

4. Cunningham, K., Kenneth, R., Koedinger, Skogsholm, A., Leber, B. (2008). An open repository and analysis tools for fine-grained longitudinal learner data. Human Computer Interaction Institute, Carnegie Mellon University. Available at: http://www.educationaldatamining.org/EDM2008/uploads/proc/16_Koedinger_45.pdf

5. Xie T., Pei, J. (2006). JMAPO: mining API usages from open source repositories. Proceedings of the International Workshop on Mining Software Repositories (MSR '06)ACM. Press, New York. Shanghai, Chinapp, 54–57. Available at: http://people.engr.ncsu.edu/txie/publications/msr06-mapo.pdf doi: 10.1145/1137983.1137997

6. Zimmermann, T. (2006). Knowledge Collaboration by Mining Software Repositories. Saarland University, Saarbrucken, Germany. Available at: http://thomas-zimmermann.com/publications/files/zimmermann-kcsd-2006.pdf

*Рекомендовано до публікації д-р техн. наук Шамша Б. В.*
*Дата надходження рукопису 20.04.2015*

**Шатовская Татьяна Борисовна**, кандидат технических наук, доцент, кафедра программной инженерии, Харьковский национальный университет радиоэлектроники, пр. Ленина, 16, г. Харьков, Украина, 61166
E-mail: shatovska@gmail.com
**Каменева Ирина Витальевна**, кандидат технических наук, доцент, кафедра программной инженерии, Харьковский Национальный университет радиоэлектроники, пр. Ленина 16, г. Харьков, Украина, 61166
E-mail: irina.kamenieva@gmail.com