УДК 330
DOI: 10.31891/2307-5740-2020-278-1-19

СИНЬКО А. І.,
ПЕЛЕЩИШИН А. М.
Національний університет "Львівська політехніка"

ЗАСТОСУВАННЯ КЛАСТЕРИЗАЦІЇ НА ОСНОВІ САМООРГАНІЗАЦІЙНИХ КАРТ КОХОНЕНА ДЛЯ РОЗПОДІЛУ КОРИСТУВАЧІВ НА ГРУПИ

В роботі наведено результати дослідження, отриманих завдяки застосуванню одного з провідних засобів аналізу даних – кластеризації. Для кластеризації даних були застосовані штучні нейронні мережі – мережа Кохонена. Для проведення дослідження обрано форум, що містить спільноти, на прикладі однієї з яких зібрані особисті дані користувачів, впорядковані за групами, відповідно за їх перебуванням на форумі, та побудована карта Кохонена для кращого представлення, розуміння та сприйняття даних. На основі отриманих даних дослідження була проведена оптимізація результатів та прогнозування. Також представлені переваги та недоліки при застосуванні даного підходу.

Ключові слова: аналіз користувачів, кластеризація, кластерний аналіз, навчання без супервізора, мережа Кохонена, форум, спільнота, мережа Інтернет.

SYNKO A.,
PELESHCHYSHYN A.
Lviv Polytechnic National University

APPLICATION OF CLUSTERIZATION ON THE BASIS OF KOHANEN'S SELF-ORGANIZING MAPS FOR DISTRIBUTION OF USERS TO GROUPS

Today the World Wide Web contains many information, data which helps to solve various issues. However, we have a problem with how this information is reliable. Therefore, the main aim of this work is the analysis of users who have published materials in virtual communities. One of the leading methods of analysis data is clustering. Cluster analysis is useful when you need to classify a large amount of information. For the clustering was selected CyberForum which have communities. Data are from the community that contains 80 users. Users divided into three groups, which are different from each other by time of stay in the forum (junior group – users who have been using the online service for less than a year; middle group – users who have been using the online service from one to five years; senior group – users who have been using the online service from five years and more). Clustering feature of a user are the number of posts he has written, experience in the field and reviews from others users. So, we have three groups of users with three characteristics for clustering. For solving the task, we chose Kohonen neural network (KNN). This method has its own advantages and disadvantages which I provided. The Python language selected for software implementation. Thousand iterative trainings selected for clustering. Because we have a lot of data for analysis, they have been taken out in a separate file. The results of study is designed map. So, we can do the next conclusions. Firstly, every user can be useful, no matter their age, job post, experience and time spent in forum or community. Secondly, for better selecting data about users' developers of software (for forums) need to encourage people, which registered on site enter more information about yourself. Thirdly, developers need to create more user selection functions (filters) for search materials.

Key words: user analysis, clustering, cluster analysis, training without supervisor, Kohonen network, forum, community, Internet.

Постановка проблеми. Глобальна мережа Інтернет на сучасному етапі розвитку людства стала всеохоплюючим явищем світового масштабу. Вона назавжди змінила життя людини, поведінку користувачів Інтернету. Сьогодні людство все частіше спілкується, проводить дослідження, шукає інформацію та підтримку щодо вирішення різних проблем в Інтернеті. Володіння достовірною інформацією про якісну оцінку користувачів (авторів), що публікують будь-які матеріали, дозволяє швидко знайти та відібрати необхідну інформацію, що постає важливим та актуальним питанням через її надмірне нагромадження у мережі Інтернет.

Аналіз останніх джерел. Одним з провідних найпоширеніших методів аналізу даних є кластеризація. Завданням якої є розбиття сукупності об'єктів на однорідні групи (класи або кластери), а метою – пошук існуючих структур. Вирішується завдання кластеризації за допомогою різноманітних методів, вибір яких повинен базуватися на дослідженні вихідного набору даних. Складністю кластеризації є необхідність її експертної оцінки.

Теоретичним аспектам застосування кластерного аналізу присвячені наукові роботи багатьох вітчизняних і закордонних вчених, зокрема Б. Еверіт, Д.С. Черезов, Т. Харріс, Р.О. Ткаченко, А.А. Барсегян, С.Л. Шульц, Л. Янг та ін. ([1–5]). Ці та інші автори сформували математичну базу для застосування кластерного аналізу в різних галузях.

Питанням якісного розподілу користувачів на групи на основі кластеризації приділяють увагу науковці, до числа яких належать В. Головко [6], Б. Сойліс [7] та ін. Незначна кількість наукових досліджень в цій галузі викликає необхідність розвивати та удосконалювати методику кластерного аналізу для якісної характеристики авторів публікацій.

Метою роботи є виявлення конкурентних переваг користувачів, що є зареєстрованими на форумі у визначеній спільноті, за допомогою самооганізаційних карт Кохонена, які використовують навчання без вчителя. І надалі, завдяки отриманій кластеризації, провести оптимізацію та прогнозування.

Виклад основного матеріалу. Сьогодні світ переповнений різною інформацією та даними – відсотками продажів, прогнозами погоди, фінансовими показниками тощо. Тому часто виникають завдання щодо аналізу даних, які насилу можна представити в математичній числовій формі. Наприклад, коли необхідно витягти дані, принципи відбору яких наведені нечітко: перевірити надійність банків або кредитоспроможність клієнтів, визначити перспективний товар, виділити користувачів за їх досвідом роботи, професійними навичками і т.п. І для того, щоб отримати максимально точні результати вирішення цих завдань потрібно застосовувати різні методи аналізу даних. Зокрема, можна використовувати штучні нейронні мережі для кластеризації даних, що, на нашу думку, є найбільш перспективним підходом.

Для розв'язку поставленої мети було обрано застосувати кластерний аналіз, який широко використовується в різних областях. Він є корисним, коли потрібно класифікувати велику кількість даних [8]. Зазвичай кластеризація є початковим етапом математичного дослідження об'єктів, за яким слідують наступні кроки, такі як оптимізація і прогнозування [9]. Одним з найбільш важливих завдань при застосуванні кластерного аналізу в цьому дослідженні є аналіз якісної оцінки користувачів, а саме: групування користувачів в однорідні класи для отримання максимально повного уявлення щодо їх стажу роботи, активності та репутації (відібрано завдяки відгукам інших користувачів) на форумі, а також і про фактори, що впливають на його поведінку. Результатом застосування кластерного аналізу є побудована карта, за якою можливо визначити якісний рівень користувача, який є учасником спільноти на форумі, що значно полегшує сприйняття даних, та надає можливість надалі висувати нові гіпотези.

Задача. На сьогоднішній день існує безліч онлайн сервісів, таких як Reddit, Stack overflow або Cyberforum, де зареєстровані користувачі можуть публікувати свої статті, наукові матеріали, виставляти пости або прямі посилання за певною тематикою, а також залишати коментарі та відгуки щодо інших робіт. Отож виникає потреба у відборі користувачів відповідно до їх знань, досвіду роботи, відгуків від інших користувачів та ін.

Для проведення кластеризації було обрано форум – Cyberforum (https://www.cyberforum.ru/), спільнота «Любители Delphi», що містить 80 користувачів.

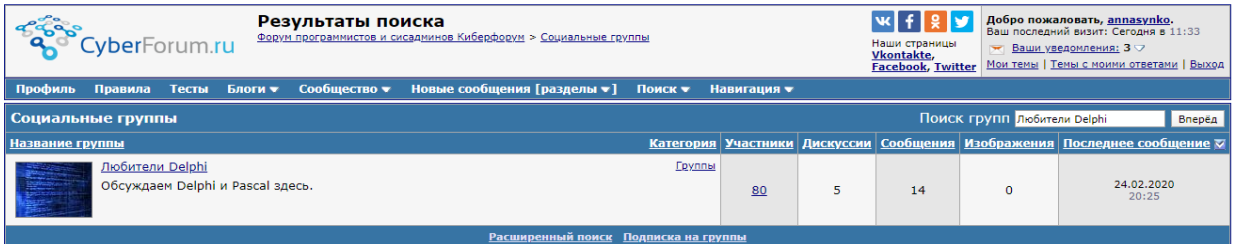

Рис. 1. Спільнота «Любители Delphi»

Дана кластеризація є описовою процедурою, вона не робить жодних статистичних висновків, адже метою кластеризації є пошук існуючих структур, але, натомість, дає можливість провести «розвідувальний» аналіз і вивчити «структуру даних» [8].

Для прикладу, наведемо інформацію про одного з користувачів, що є учасником цієї спільноти:

Рис. 2. Інформація про обраного користувача зі спільноти

На основі цих даних можна провести кластеризацію. Оскільки сама кластеризація не дає конкретних результатів аналізу, тому для отримання ефекту необхідно виконати змістовну інтерпретацію кожного кластера. Отже, спираючись на дані про перебування користувачів на форумі поділимо їх на три групи:

– ті, хто перебуває менше року на онлайн-сервісі (junior);

– ті, хто перебуває від одного до п'яти років на онлайн-сервісі (middle);

– ті, хто перебуває від п'яти років на онлайн-сервісі (senior).

На рис. 2 бачимо, що користувач перебуває на форумі з лютого 2015 року, тому можемо віднести його до третьої групи – senior.

Кластеризація даних поставленої задачі буде реалізовуватися в такі етапи:

А. Виділення характеристик (відбір властивостей, які характеризують обрані об'єкти. Отримані дані необхідно нормалізувати. Далі всі об'єкти надаються у вигляді характеристичних векторів (рис. 3), що надає можливість надалі ототожнювати об'єкт з його характеристичним вектором.

Б. Визначення метрики (вибір метрики, за якої визначається близькість об'єктів). Метрика вибирається залежно від простору, в якому розташовані об'єкти; неявних характеристик кластерів. Зазвичай використовують класичну евклідову метрику – формула 1.

$$d_2(x_i, x_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2} = \left\| x_i - x_j \right\|_2 . \quad (1)$$

В. Представлення результатів в зручному для обробки вигляді для подальшої оцінки якості кластеризації (було застосовано представлення кластерів набором характерних точок).

Характеристиками користувачів для виявлення внутрішніх взаємозв'язків, залежностей, закономірностей, що існують між об'єктами, є:

- кількість постів (повідомлень) – активність; оцінюємо від 1 до 5;

- досвід роботи в галузі. Оцінюємо від 1 до 10;

- відгуки від інших користувачів (репутація). Оцінюємо від 1 до 6 (оцінка 1– 1-2 бали, оцінка 2 – 3-5 балів і т.д.);

Виходячи з вищесказаного обираємо нейромережу, яка реалізується методом навчання без супервізора [10].

Навчання без вчителя – один зі способів машинного навчання, при вирішенні яких обрана система спонтанно навчається виконувати поставлене перед нею завдання, без втручання з боку експериментатора. Зазвичай, це підходить тільки для задач, в яких заздалегідь відомий опис множини об'єктів (навчальна вибірка), і потрібно виявити внутрішні закономірності, взаємозв'язки, залежності, що існують між об'єктами. Засобами розв'язання таких задач є глибинна мережа переконань, графові алгоритми кластеризації, кластеризація методом k-середніх, нейронна мережа Кохонена. Для розв'язку поставленого завдання було обрано нейронну мережу Кохонена, яка має свої переваги, що наведені нижче.

Програмна реалізація. Програмне забезпечення, що дозволяє працювати з картами Кохонена, зараз представлено безліччю інструментів. Це можуть бути як інструменти, що включають тільки реалізацію методу самоорганізаційних карт, так і нейропакет з цілим набором структур нейронних мереж, серед яких і карти Кохонена.

До інструментарію, що включає реалізацію методу Кохонена відносяться NeuroShell, Statistica, MATLAB Neural Network Toolbox, NeuroScalp, SoMine, Deductor тощо. Для розв'язання поставленого завдання обрано мову програмування Python і застосовано її вбудовані функції, команди.

Через те, що об'єктів є багато (перша група – 22 користувача, друга група – 27, третя група – 31) всі їх дані були занесені в окремий файл (рис. 3). Отже, в нас є три групи користувачів (всього 80 об'єктів), кожен з яких має три характеристики. Для навчання було обрано 1000 ітерацій. Карта має розмір 7×7 (рис. 4).

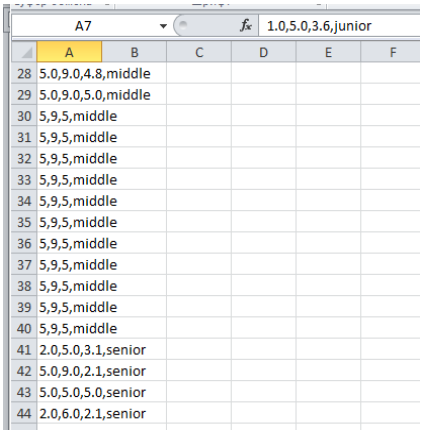

Рис. 3. Дані, що є впорядкованими відповідно до груп користувачів

Процес навчання карти Кохонена відбувається в такі етапи: етап впорядкування векторів вагових коефіцієнтів в просторі ознак і етап підстроювання.

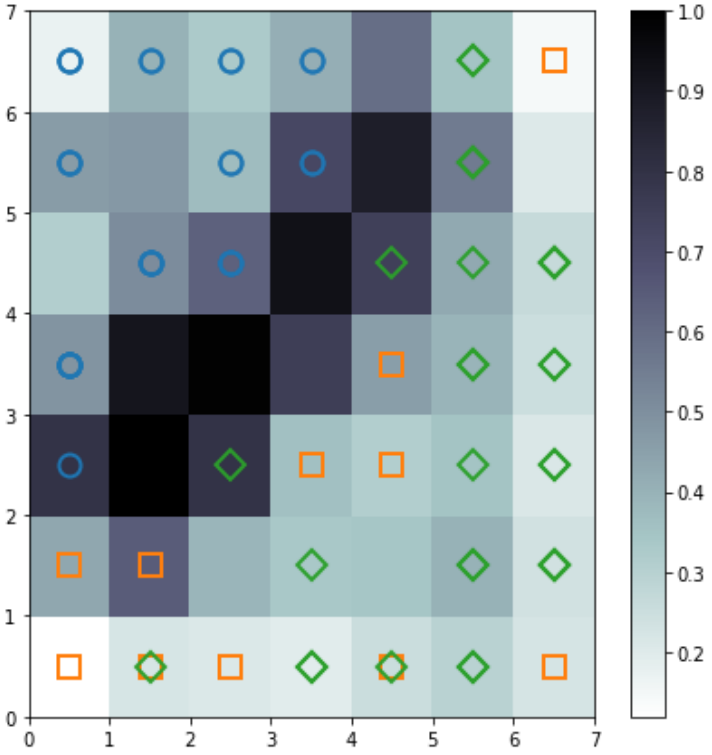

Рис. 4. Розв'язок задачі застосовуючи карту Кохонена, що пройшла навчання без супервізора

Як видно на рис. 4, кожна група користувачів має свій колір: junior – синій; middle – зелений; senior – коричневий. Також була побудована шкала, що відображає відстані між об'єктами (чим темніше зафарбування, тим більше відстань).

Однак, слід наголосити, що цей експеримент не є самоціллю, адже кінцевою метою кластеризації є отримання змістовних відомостей про структуру досліджуваних даних. Отримані результати вимагають подальшої інтерпретації, дослідження і вивчення властивостей і характеристик об'єктів для можливості точного опису сформованих кластерів [9].

Для кращого розуміння розв'язку поставленої задачі – виявлення конкурентних переваг користувачів, що є зареєстрованими на форумі, – за допомогою самоорганізаційних карт Кохонена побудуємо схему (рис. 5 та рис. 6).

Рис. 5. Загальна схема розв'язку поставленої задачі

Рис. 6. Проходження навчання без супервізора

Перевагами застосування даного методу для вирішення поставленої задачі є:
– стійкість до зашумлення даних;
– некероване навчання;
– можливість візуалізації (побудована карта);
– швидке навчання;
– можливість спрощення багатомірної структури.

Як і будь-кий метод може мати свої недоліки, обрана система має свої:

– вибір коефіцієнта навчання (впливає як на стійкість одержуваного рішення, так і на швидкість навчання);

– вибір початкових значень нейронів і векторів вагових коефіцієнтів (якщо початкові значення обрані невдало, тобто, наприклад, розташовані далеко від пропонованих вхідних векторів, то нейрон не опиниться переможцем ні за яких вхідних сигналів, а, отже, не навчиться);

– рандомізація ваг (рандомізація ваг прошарку Кохонена може породити серйозні проблеми при проходженні навчання, так як в результаті цієї процедури вагові вектори рівномірно розподіляються по поверхні гіперсфери. Зазвичай, вхідні вектори нерівномірно розподілені і групуються на відносно малій частині поверхні гіперсфери. Тому більшість вагових векторів виявляться настільки віддаленими від будь-якого вхідного вектора, що не активовані і стануть марними. Більше того, активованих нейронів, які залишилися, може виявитися занадто мало, щоб розбити близько розташовані вхідні вектори на кластери);

– вибір параметра відстані (якщо обраний на початку параметр є малим або дуже швидко зменшується, то далеко розташовані один від одного нейрони не зможуть впливати один на одного. Хоча дві частини в такій карті налаштовуються правильно, загальна карта буде мати топологічний дефект, рис. 7).

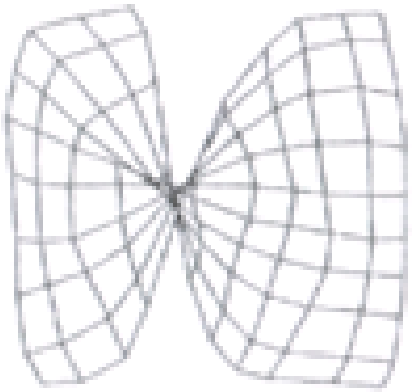

Рис. 7. Топологічний дефект карти

Отже, провівши наукове дослідження, врахувавши всі недоліки та переваги обраного методу, отримали наступний результат. По-перше, кожен користувач, що перебуває у спільноті, може бути корисним (розміщати актуальні та достовірні матеріали, публікації не зважаючи на його вік, посаду, досвід роботи тощо), тому що група користувачів junior, ті що перебувають у спільноті менше року, це можуть бути молоді люди, які не могли зареєструватися на сайті десять років тому, мають доволі високі показники відповідно до характеристик, за якими проводилось дослідження. По-друге, щоб краще відбирати дані про користувачів, які роблять публікації, потрібно розробникам заохочувати користувачів вводити якомога більше інформації про себе, або зробити ці пункти обов'язковими полями для заповнення при реєстрації. Так як, зрозуміло, що може бути похибка, адже обрано замало характеристик для дослідження даної галузі (чим більше даних тим менше похибка). Що спонукає розробників замислитися над цим питанням. По-третє, даний метод відбору користувачів має зацікавити аналітиків, розробників створювати додаткові функції по відбору користувачів, щоб шукаючи потрібну інформацію на форумі або у спільноті, можна було обрати «найкориснішого» автора матеріалів в обраній галузі.

Висновки. Отже, дивлячись на розв'язок завдання, можна зробити висновок: незалежно від того, як довго користувач перебуває на будь-якому онлайн ресурсі (і, обов'язково, є зареєстрованим на ньому), він може бути не менш корисним для суспільства, навіть незважаючи на його вік, посаду, досвід роботи тощо. Звичайно, це важливі чинники, якими не можна нехтувати, але тут має силу ще такий фактор, як швидко людина знаходить, сприймає та опановує нову інформацію. Наприклад, якщо раніше були потрібні роки для вивчення тієї чи іншої дисципліни, то тепер це стало доступнішим завдяки новим технологіям і попереднім відкриттям. Також, якщо раніше, щоб бути фахівцем у сфері програмування, потрібні були роки, то зараз кілька місяців, так як кожен має доступ до комп'ютера та мережі Інтернет, де може знайти теорію та практикуватися.

Звісно мережа Інтернет має і свої недоліки, тому що існує багато інформації, яка не є достовірною або, навіть, хибною. Для цього і була створена класифікація користувачів за їхніми характеристиками, щоб відібрати якомога якісну, достовірну інформацію.

Через те, що кластеризація є одним з провідних методів аналізу даних, було обрано один з її підходів – нейронну мережу Кохонена. Для чіткого уявлення щодо послідовності кроків, яка пройшла система для досягнення поставленого завдання, була наведена схема – рис. 5 та рис. 6. Також були описані переваги та недоліки при застосуванні обраного підходу, якими не слід нехтувати при проведенні інших подібних досліджень.

Обраний напрям дослідження є актуальним та потребує подальших напрацювань, адже проблема надлишкової інформації, причому не завжди достовірної, є сьогоденною і важливою для кожного, хто

користується Інтернетом для пошуку будь-яких даних. Отже, питання відбору якісної, достовірної інформації, яку публікують користувачі, залишається не повністю розкритим, адже є й інші чинники, за якими можна проводити аналіз даних (наприклад, зробити обов'язковим для користувачів посилання на літературні джерела – звідки вони взяли цю інформацію). Також це можуть бути такі характеристики: володіння іноземними мовами, посада, яку займає користувач тощо). Щоб втілити ці всі характеристики для подальшого аналізу, потрібно розробникам таких спільнот та форумів заохочувати користувачів вводити якомога більше інформації про себе, або зробити ці пункти обов'язковими полями для заповнення під час реєстрації.

Література

1. Everitt B., Landau S., Leese M., Stahl D. Cluster Analysis. Wiley, 2010. 346 p.
2. Черезов Д. С. Обзор основных методов классификации и кластеризации данных / Д. С. Черезов, Н. А.Тюкачев // Вестник ВГУ, серия: системный анализ и информационные технологии. – 2009. – № 2. – С. 25–29.
3. Ткаченко Р. О. Нейромережеві засоби штучного інтелекту : навчальний посібник / Ткаченко Р. О., Ткаченко П. Р., Ізонін І. В. – Львів : Видавництво Львівської політехніки, 2017. – 204 с.
4. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. / [Барсегян.А. А., Куприянов М. С., Степаненко В. В., Холод И. И.]. – 2-е изд. – СПб : БХВ-Петербург, 2007. – 384 с.
5. Wang J. H., Rau J. D., and Liu W. J. Two-stage clustering via neural networks. IEEE Transactions on Neural Networks, 2003, Vol. 14, pp. 606–615.
6. Головко В.А. Нейронные сети: обучение, организация и применение. Кн. 4 : учеб. пособие для вузов / общ. ред. А.И. Галушкина. – Москва : ИПРЖР, 2001. – 256 с. – (Нейрокомпьютеры и их применение).
7. Брайан Соліс. Роль сучасних соціальних мереж в соціумі та політичних технологіях / Брайан Соліс – Москва : Директ-Медиа, 2012.
8. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям : учеб. пособие / Паклин Н.Б., Орешков В.И. – 2-е изд., перераб. и доп. – СПб : Питер, 2013. – 704 с.
9. Дебок Г. Анализ финансовых данных с помощью самоорганизующихся карт / Г. Дебок, Т. Кохонен ; пер. с англ. – М. : Альпина, 2001. – 317 с.
10. Кохонен Т. Самоорганизующиеся карты / Т. Кохонен ; пер. с англ. В. Агеев ; под ред. Ю. Тюменцева. – М. : Бином, 2008. – 656 с.

References

1. Everitt B., Landau S., Leese M., Stahl D. Cluster Analysis. Wiley, 2010. 346 p.
2. Cherezov D. S. Obzor osnovnyh metodov klassifikacii i klasterizacii dannyh / D. S. Cherezov, N. A.Tyukachev // Vestnik VGU, seriya: sistemnyj analiz i informacionnye tehnologii. – 2009. – № 2. – S. 25–29.
3. Tkachenko R. O. Neiromerezhevi zasoby shtuchnoho intelektu : navchalnyi posibnyk / Tkachenko R. O., Tkachenko P. R., Izonin I. V. – Lviv : Vydavnytstvo Lvivskoi politekhniky, 2017. – 204 s.
4. Tehnologii analiza dannyh: Data Mining, Visual Mining, Text Mining, OLAP. / [Barsegyan.A. A., Kupriyanov M. S., Stepanenko V. V., Holod I. I.]. – 2-e izd. – SPb : BHV-Peterburg, 2007. – 384 s.
5. Wang J. H., Rau J. D., and Liu W. J. Two-stage clustering via neural networks. IEEE Transactions on Neural Networks, 2003, Vol. 14, pp. 606–615.
6. Golovko V.A. Nejronnye seti: obuchenie, organizaciya i primenenie. Kn. 4 : ucheb. posobie dlya vuzov / obsh. red. A.I. Galushkina. – Moskva : IPRZhR, 2001. – 256 s. – (Nejrokompyutery i ih primenenie).
7. Braian Solis. Rol suchasnykh sotsialnykh merezh v sotsiumi ta politychnykh tekhnolohiiakh / Braian Solis – Moskva : Dyrekt-Medya, 2012.
8. Paklin N.B. Biznes-analitika: ot dannyh k znaniyam : ucheb. posobie / Paklin N.B., Oreshkov V.I. – 2-e izd., pererab. i dop. – SPb : Piter, 2013. – 704 s.
9. Debok G. Analiz finansovyh dannyh s pomoshyu samoorganizuyushihsya kart / G. Debok, T. Kohonen ; per. s angl. – M. : Alpina, 2001. – 317 s.
10. Kohonen T. Samoorganizuyushiesya karty / T. Kohonen ; per. s angl. V. Ageev ; pod red. Yu. Tyumenceva. – M. : Binom, 2008. – 656 s.

Рецензія/Peer review : 12.01.2020 Надрукована/Printed : 10.03.2020
Рецензент: д. е. н., проф. Войнаренко М. П.