УДК 004.891.2

РОЗРОБКА ЕЛЕМЕНТІВ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ НА БАЗІ ВЕБ-АНАЛІТИКИ

Канакін А.Е., Скарга-Бандурова І.С., Щербакова М.Є.

ELEMENTS OF RECOMMENDATION SYSTEM BASED ON WEB-ANALYTICS

Kanakin A.E., Skarga-Bandurova I.S., Scherbakova M.E.

Стаття присвячена аналізу методів й інструментів розробки рекомендаційної системи, здатної виконувати збір, аналіз, представлення й інтерпретацію інформації про відвідувачів веб-сайтів з метою поліпшення їх обслуговування. Розглянуті сервіси веб-аналітики Google Analytics, їх переваги і недоліки. Запропоновано варіант власної системи, що побудована з використанням колаборативної фільтрації та для пошуку подібних вподобань користувачів використовує косинусну схожість.

Ключові слова: *веб-аналітика, рекомендаційна система, колаборативна фільтрація, онлайн-магазин.*

Вступ. Сучасний бізнес неможливо уявити без інтернету. З активним розвитком веб-ресурсів виникли питання, пов'язані з їх функціональними можливостями для проведення моніторингу відвідуваності сайтів, оцінкою економічної ефективності рекламних інтернет-кампаній та оптимізацією структури і вмісту сайту.

Рішенням вищевказаних проблем займається веб-аналітика. Сучасна веб-аналітика дозволяє збирати і аналізувати інформацію про відвідувачів інтернет ресурсів. Встановлюючи лічильники аналітичних систем на свої сайти, рекламодавці отримують можливість відстежувати клієнтів на всіх етапах, від кліка по рекламному оголошенню до оплати замовлення в інтернет-магазині. Основним завданням веб-аналітики є моніторинг відвідуваності веб-сайтів, на підставі якого визначається аудиторія сайту і вивчається поведінка відвідувачів для прийняття рішень щодо розвитку і розширення функціональних можливостей веб-ресурсу. Загалом, веб-аналітика дозволяє не тільки працювати над поліпшенням сайтів, але й проводити роботи по оптимізації бюджету на онлайн-просування.

Одним з найбільш поширених засобів відстеження поведінки відвідувачів веб-сайтів є сервіси Google analytics. Це система вимірювання, збору, аналізу, представлення й інтерпретації інформації про відвідувачів веб-сайтів з метою їх поліпшення і оптимізації. Іншим популярним засобом, що активно використовує веб-аналітику, є рекомендаційні системи, які не тільки відслідковують поведінку користувачів інтернет-сервісів, але й надають користувачам оцінки щодо переваг того чи іншого об'єкта. Об'єктами рекомендацій можуть служити товари в інтернет-магазині, набір розділів веб-сайту, медіа-контент, інші користувачі веб-сервісу. На основі сформованих рекомендаційною системою переваг, поведінка веб-сервісу для кожного конкретного користувача може змінюватися, надаючи персоніфікований контент. На даний момент, рекомендаційні системи знаходяться в стадії активного розвитку і вимагають розробки технологій, здатних надавати найкращі, в певному сенсі, поради, адаптовані під конкретного користувача. Така задача вимагає вирішення багатьох питань, одним з яких є пошук найкращих моделей, що описують схожість.

Метою статті є аналіз методів й інструментів для розробки рекомендаційної системи, та представлення варіанту власної системи, що побудована з використанням колаборативної фільтрації.

Основна частина. Найбільш розповсюдженою системою веб-аналітики є Google Analytics (GA), що являє собою безкоштовний інтернет-сервіс для створення детальної статистики відвідувачів веб-сайтів. Статистика відвідувачів виконується за рахунок JS-код на сторінках свого сайту. За умови дозволеного виконання Javascript, код відстеження (JS-код) спрацьовує при кожному відкритті сторінки користувачем. GA - це не просто лічильник відвідувань, це повноцінний інструмент аналізу ефективності роботи сайту компанії та її проведених маркетингових заходів. Google Analytics дозволяє не тільки відслідковувати джерела відвідувачів сайту, але й аналізувати їх ефективність. У GA створена система вибору готових статистичних звітів, які

базуються на наступних функціональних можливостях [1]:

- відстеження цілей;
- інтеграція Google Analytics з Google AdWords і Google AdSense;
- відстеження продажів інтернет-магазинів;
- відстеження мобільних пристроїв;
- відстеження внутрішнього пошуку по сайту;
- порівняння показників;
- відстеження використання Flash, Ajax і відео;
- розширена сегментація в Google Analytics;
- призначені для користувача звіти;
- експорт даних у формати Excel, CSV, PDF;
- відправка звітів по електронній пошті;
- API для розробників.

Комбінування різних функціональних можливостей дозволяє формувати звіти, що найбільш повно відповідають запитам. Вивчення поведінки користувачів на сайті, оцінка юзабіліті й аналіз відвідуваності сайту будується на звітах, сформованих за певними критеріями. Сервіс інтегрований з Google AdWords. Особливістю сервісу є те, що веб-майстер може оптимізувати рекламні та маркетингові кампанії Google AdWords за допомогою аналізу даних, отриманих за допомогою сервісу Google Analytics, зібрати інформацію про шляхи приходу відвідувачів, час перебування на сайті, географічне місце розташування відвідувачів. Використовуючи цей інструмент, маркетологи можуть визначати, яка з рекламних кампаній є успішною, і знаходити нові джерела та цільові аудиторії [2].

Одним з ключових аналітичних інструментів Google Analytics є візуалізація даних [1]. Звіт по візуалізації являє собою графічне представлення показників обраних користувачем і дозволяє порівнювати обсяги трафіку з різних джерел, вивчати структуру трафіку і вимірювати ефективність сайту.

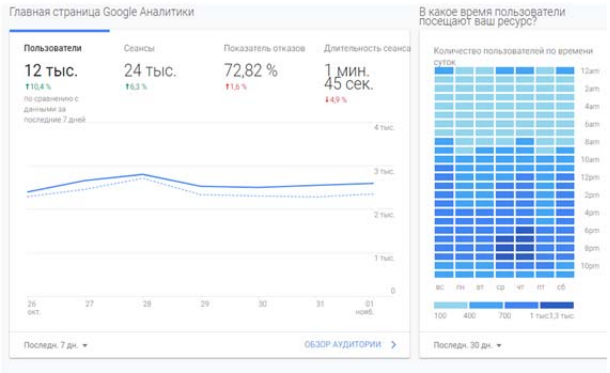

Рис. 1. Візуалізація даних в Google Analytics

Візуалізація переходів дає можливість покращувати конверсію і визначати, що з контенту подобається відвідувачам, а що ні [3]. Окрім зазначеного, Google Analytics дозволяє аналізувати поведінку і характеристику споживачів на сайті з деталізацією кожного візиту, пошукового запиту і джерела трафіку [4]. Система фільтрів зведень і звітів дозволяє сегментувати відвідувачів за різними показниками та їх сполученням.

До основних недоліків Google Anlatics можна віднести:

– обмеження по глибині доступу до даних (немає можливості стежити за кожним конкретним візитом або кожним конкретним кліком);

– інформація, що збирається Google Analytics, належить не користувачу;

– неможливість вилучити з бази даних власні статистичні дані, наприклад, якщо потрібно використовувати іншу аналітичну систему;

– аналізуючи отриману інформацію, важко сформувати, яка конструкція менш відштовхує відвідувачів сайту, і змушує поглибити вивчення контенту, що повинно бути написано на баннері, щоб на нього частіше натискали, де повинна бути контактна форма, щоб вона частіше заповнювалася, який дисконт потрібно надати, для більшої конверсії, і багато іншого. Для проведення спліт-тестування необхідно використовувати ще один інструмент від Google - Оптимізатор веб-сайтів Google.

Отже, для вирішення задачі аналізу поведінки й характеристик споживачів на сайті, має сенс використовувати спеціалізовані сервіси – рекомендаційні системи.

Можна виділити три основні підходи до побудови рекомендаційних систем: на підставі аналізу змісту (content-based); колаборативна фільтрація (collaborative filtering); гібридний підхід.

Підхід на підставі аналізу змісту передбачає, що про користувачів і про рекомендовані об'єкти є досить багато інформації. Наприклад, всі користувачі заповнюють анкету, в якій вказують свою соціально-демографічну інформацію, інтереси, і т.д. Про товари з інтернет-магазину можуть бути відомі їх опис, призначення, цінова категорія, бренд, та інші характеристики. Маючи історію взаємодії користувачів і об'єктів на сервісі можна побудувати навчальну вибірку і провести прогнозування за добре відомими прикладами.

На практиці, використання такого підходу сильно обмежене, оскільки збір описової інформації про користувачів і об'єкти є дуже дорогою процедурою, яку часто неможливо організувати, не знижуючи якість використання сервісу, що робить рекомендаційну систему, побудовану за цим підходом, невиправдано дорогою.

Колаборативна фільтрація. Колаборативна фільтрація зазвичай використовується за умов, коли рекомендаційна система не володіє будь-якою інформацією про користувачів і об'єкти (або не використовує) і будує прогноз виключно на підставі взаємодії користувачів з об'єктами.

Нехай U - множина користувачів (users), I - множина об'єктів (items), інформація про відомі переваги, представлена у вигляді набору трійок:

$$D = \{(u,i,r_{ui})\}, (u,i) \in R,$$

де $r_{ui} \in R$ - речова ступінь переваги об'єкта $i \in I$ користувачем $u \in U$; $R \subseteq U \times I$ - множина пар (користувач, об'єкт), про які відома ступінь переваги.

Для подальшої зручності, введемо також позначення: $R(u) = \{i : (u,i) \in R\}$ - множина об'єктів, суміжних з користувачем *u,* аналогічно: $R(i) = \{u : (u,i) \in R\}$.

За відомою інформацією D потрібно побудувати прогноз переваги $r_{ui} \approx r_{ui}$ для нових пар $(u,i) \in R$.

Будемо називати матрицею оцінок - матрицю $R \in (R \cup \varnothing)\ |U| \times |I|$, рядки якої відповідають користувачам, стовпці - об'єктам, а елементи приймають значення r_{ui}, якщо $(u,i) \in R$, та $\varnothing$ в протилежному випадку.

Тоді, на завдання колаборативної фільтрації можна дивитися як на задачу заповнення пропущених значень в матриці.

Крім передбачення значень переваги, на практиці можуть бути цікаві такі завдання:

- побудова списку рекомендацій з об'єктів, на які не відома ступінь переваги (нові для користувача):

Recommend $K(u) = \text{Top}_K \max_i r_{ui} \rightarrow \{(i_1, r_{ui1}), (i_2, r_{ui2}), ..., (i_K, r_{uiK})\}$;

- визначення ступеню схожості об'єктів і побудова списків найбільш схожих:

Similar K $(i) \rightarrow \{(i_1, s_{i1}), (i_2, s_{i2}), ..., (i_K, s_{iK})\}$,

де s_{ij} - ступінь схожості між двома об'єктами;

- обґрунтування списку рекомендацій: деяке людино-зрозуміле пояснення, чому користувачеві *u* необхідно порекомендувати об'єкт *i.*

Підходи до вирішення завдання колаборативної фільтрації умовно можна розділити на дві великі групи:

1. Засновані на евристиках (memory / heuristic-based);
2. Засновані на побудові моделі переваги (model-based).

До першої групи методів (memory-based) відносяться алгоритми, що виражають припущення значення безпосередньо через елементи матриці оцінок. Прикладом memory-based алгоритму колаборативної фільтрації є зважування переваги по користувачах (user-based):

$$\hat{r}_{ui} = \overline{r}_u + \frac{1}{\sum_{u' \in R(i)} |sim(u,u')|} \sum sim(u,u')(r_{u',i} - \overline{r}_{u'}),$$

і по об'єктах (item-based):

$$\hat{r}_{ui} = \overline{r}_i + \frac{1}{\sum_{i' \in R(i)} |sim(i,i')|} \sum sim(i,i')(r_{u,i'} - \overline{r}_{i'}),$$

де $\overline{r}_u = \frac{1}{|R(u)|} \sum_{i \in R(u)} r_{ui}$, $\overline{r}_i = \frac{1}{|R(i)|} \sum_{u \in R(i)} r_{ui}$ – середні значення переваг по користувачах і об'єктам, sim *(u, u')*, sim *(i, i')* – відповідно, наперед задані метрики схожості користувачів і об'єктів.

Міра схожості sim(u, u^0) (і аналогічна для об'єктів) обчислюється по матриці оцінок *R,* або з використанням додаткової інформації про користувачів (об'єкти). Найбільш вживаними є прості метрики схожості, такі як кореляція Пірсона:

$$sim(u,u') = \frac{\sum_{i \in R(u) \cap R(u')} (r_{u,i} - \overline{r}_u)(r_{u',i} - \overline{r}_{u'})}{\sqrt{\sum_{i \in R(u) \cap R(u')} (r_{u,i} - \overline{r}_u)^2 \sum_{i \in R(u) \cap R(u')} (r_{u',i} - \overline{r}_{u'})^2}}$$

і косинусна відстань відповідних рядків (стовпців) матриці оцінок:

$$sim(u,u') = \frac{\sum_{i \in R(u) \cap R(u')} r_{u,i} \cdot r_{u',i}}{\sqrt{\sum_{i \in R(u)} r_{u,i}^2} \cdot \sqrt{\sum_{i \in R(u')} r_{u',i}^2}}.$$

Найбільш використовуваними і реалізованими у вигляді бібліотек з відкритим вихідним кодом є memory-based алгоритми. Ці алгоритми є корисними для одноразового обчислення рекомендацій на розподіленому кластері, вони добре працюють з MapReduce, проте погано підходять для оперативного оновлення рекомендацій. Налаштування схожості для завдання з конкретної предметної області все ще представляє собою складну задачу.

Одна з найбільш серйозних проблем memory-based методів - неадекватність передбачень в умовах сильної розрідженості матриці оцінок R (в сенсі пропущених значень), що призводить до неможливості підрахунку метрик схожості в разі, якщо множина $R(u) \bigcap R(u^0) = \varnothing$. Сильна розрідженість матриці оцінок може бути наслідком проблеми холодного старту, проте в деяких областях вона є сильно розрідженою завжди і не може стати досить щільною (наприклад, в разі інтернет-магазинів, користувач залишає інформацію про переваги в середньому лише по 2-5 об'єктіам).

Алгоритми з другої групи (model-based) направлені на вибір функції $r(u, i; \theta)$ з деякого сімейства параметризованих моделей $\theta \in \Theta$.

order_	order_id	product_id	name	model	quantity	price	total	tax	reward
68	47	113	Сет канна	Сет	1	620.0000	620.0000	0.0000	4
67	46	151	Сет кумамото	Сет	1	530.0000	530.0000	0.0000	5
54	40	128	Филадельфия	Маки-суши	1	140.0000	140.0000	0.0000	3
55	40	151	Сет кумамото	Сет	1	530.0000	530.0000	0.0000	4
56	41	77	Кадзура	маки- суши	1	120.0000	120.0000	0.0000	3
57	42	108	Салат хияши с кальмаром и тигровой креветкой	Салат	1	90.0000	90.0000	0.0000	5
58	42	116	Сет орхидея	Сет	1	665.0000	665.0000	0.0000	4
59	43	62	Банзай	горячие роллы	1	150.0000	150.0000	0.0000	5
60	43	130	Фудзияма	Маки-суши	1	115.0000	115.0000	0.0000	4
61	44	118	Сет ханами	Сет	1	300.0000	300.0000	0.0000	3
62	45	147	Ролл с красной икрой	маки- суши	1	130.0000	130.0000	0.0000	5
63	45	81	Канада	маки- суши	1	170.0000	170.0000	0.0000	4
64	45	82	Кани-унаги	маки- суши	1	135.0000	135.0000	0.0000	5
65	45	108	Салат хияши с кальмаром и тигровой креветкой	Салат	1	90.0000	90.0000	0.0000	2
66	45	73	Ика темпура	маки- суши	1	130.0000	130.0000	0.0000	4
69	47	151	Сет кумамото	Сет	1	530.0000	530.0000	0.0000	5
70	48	115	Сет купидон	Сет	1	480.0000	480.0000	0.0000	4
71	49	70	дзен	маки- суши	1	165.0000	165.0000	0.0000	5
72	49	130	Фудзияма	Маки-суши	1	115.0000	115.0000	0.0000	5
73	49	81	Канада	маки- суши	1	170.0000	170.0000	0.0000	5
74	49	62	Банзай	горячие роллы	1	150.0000	150.0000	0.0000	5
75	50	116	Сет орхидея	Сет	1	565.0000	565.0000	0.0000	4
76	51	108	Салат хияши с кальмаром и тигровой креветкой	Салат	1	90.0000	90.0000	0.0000	5

Рис. 2. Дані про замовлення

Вибір може відбуватися, наприклад, шляхом мінімізації регуляризованого емпіричного ризику:

$$L(\theta) = \sum_{(u,i)\in R} l(\hat{r}(u,i;\theta), r_{ui}) + \lambda F(\theta) \to \min_{\theta\in\Theta},$$

де l – функція втрат регресії, λ – сила регуляризації, $F(\theta)$ – функція регуляризатора на множині параметрів Θ.

За результатами аналізу алгоритмів колаборативної фільтрації було розроблено рекомендаційну систему по схожості користувачів, що визначається з використанням косинусної схожості по оцінкам продукції. Рекомендаційна система розробляється для веб-сайту з продажу та доставки їжі.

Для розробки алгоритму потрібні данні про замовленя. Ці дані беруться с бази даних SQL *«oc_order_product»* (рис. 2), де поле «order_id» - це ідентифікатор користувача, що зберігає такі дані як ім'я, телефон, адрес и т.д.; *«product_id»* - продукт з сайту, де поле *«name»* його назва і поле *«reward»* - оцінка, яку поставив користувач цьому продукту. Усі описані поля потрібні для розробки рекомендаційною системи для інтернет магазину.

Для рекомендації користувачеві №1 будь-якого продукту, вибирати потрібно з продуктів, які подобаються якимось користувачам № 2, 3, 4 і так далі, які найбільш схожі за своїми оцінками на користувача №1.

Потрібно отримати чисельне вираження схожості користувачів. Оцінки, виставлені окремо взятим користувачем, складають вектор в M-вимірному просторі продуктів.

Косинусна міра для двох векторів - це косинус кута між ними. Косинус кута між двома векторами - це їх скалярний добуток, поділений на довжину кожного з двох векторів:

$$\cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|_2 \times \|\vec{y}\|_2}.$$

Для розрахунку косинусної міри використаємо *«poruct_id»* та *«reward»*. З отриманої матриці переваг користувачів можна легко визначати, наскільки два користувача схожі один на одного за своїм вибором. Далі необхідно реалізувати алгоритм колаборативної фільтрації. Вибрати користувачів, смаки яких найбільше схожі на смаки розглянутого. Для цього для кожного з користувачів потрібно обчислити обрану міру (в нашому випадку косинусну) щодо розглянутого користувача, і вибрати найбільших. Для активного користувача отримуємо дані (рис.3).

user_id:	12	24	35	45
текущий	0,7897	0	0,0789	0,879
all				1,7476

Рис. 3. Дані про схожість користувачів

Далі для кожного з користувачів необхідно помножити його оцінки на обчислену величину, так оцінки більш схожих користувачів будуть сильніше впливати на підсумкову позицію продукту. Також, для кожного з продуктів необхідно порахувати суму каліброваних оцінок найбільш близьких користувачів і отриману суму розділити на суму заходів обраних користувачів.

	product_1	product_2	product_3	product_4	product_5
user_12	2,127	1,892	0	0,789	1,545
user_24	0	0	0	0,48	0,245
user_35	0	0,0158	0,01	0	0
user_45	0,2578	0,0078	0	0	0,0009
all	2,3848	1,9156	0,01	1,269	1,7909
getresult	3,4789	2,5478	0,5788	1,9878	2,4547

Рис. 4. Коефіцієнт популярності товару

Формально, цей крок може бути представлений як розрахунок

$$r_{u,i} = k \sum_{u' \in U} sim(u,u')r_{u',i},$$

де функція *sim* – обрана міра схожості двох користувачів, U – множина користувачів, r – виставлена оцінка, k – нормувальний коефіцієнт: $k = 1/\sum_{u' \in U} |sim(u,u')|$.

Результат сумації представлено на рис. 4 в рядку «all», підсумкове значення в рядку «getresult».

Висновок. За результатами проведеного аналізу з'ясовано, що широко поширений сервіс веб-аналітики Google Analytics має певні недоліки для розгортання повноцінної рекомендаційної системи. З огляду на це, у роботі запропоновано елементи власної системи, що побудована з використанням колаборативної фільтрації та для пошуку подібних вподобань користувачів використовує косинусну схожість. Рекомендаційна система буде використана на сайті з продажу та доставки їжі.

Література

1. Analytical tools Google Analytics [Electronic resource]. Access mode: http://www.google.by/intl/ru_ALL/analytics/features/analysis–tools.html.
2. Analytics online store for 5 khvilin: yak trimati sale pid control [Electronic resource]. Access mode: https://www.ecwid.ru/blog/customize-dashboard-google-analytics-and-yandex-metrika.html
3. Web analytics of corporate level Google Analytics [Electronic resource]. Access mode: http://www.google.com/intl/ru_ALL/analytics/index.html.
4. Visualization of user paths to the site [Electronic resource]. Access mode: https://support.google.com/analytics/answer/1709395?hl=en&topic=1709360&ctx=topic.
5. G. Adomavicius and O. Tuzhilin. At the top of a new generation of systems, I recommend it: an eye on the most important and the most elevated. Knowledge and Engagement, IEEE Transaction, 17, 2005.
6. H. Adomavichyus that O. Tuzhilin. Context-based system recommendations. At the post office of the system "Recommendations". Springer, 2011.
7. J. Bennett and S. Lanning. Netflix prize. At the Material KDD Cup and Maysterni, volume 2007.

References

1. Analytical tools Google Analytics [Electronic resource]. Access mode: http://www.google.by/intl/ru_ALL/analytics/features/analysis–tools.html.
2. Analytics online store for 5 khvilin: yak trimati sale pid control [Electronic resource]. Access mode: https://www.ecwid.ru/blog/customize-dashboard-google-analytics-and-yandex-metrika.html
3. Web analytics of corporate level Google Analytics [Electronic resource]. Access mode: http://www.google.com/intl/ru_ALL/analytics/index.html.
4. Visualization of user paths to the site [Electronic resource]. Access mode: https://support.google.com/analytics/answer/1709395?hl=en&topic=1709360&ctx=topic.
5. G. Adomavicius and O. Tuzhilin. At the top of a new generation of systems, I recommend it: an eye on the most important and the most elevated. Knowledge and Engagement, IEEE Transaction, 17, 2005.
6. H. Adomavichyus that O. Tuzhilin. Context-based system recommendations. At the post office of the system "Recommendations". Springer, 2011.
7. J. Bennett and S. Lanning. Netflix prize. At the Material KDD Cup and Maysterni, volume 2007.

Канакин А.Е., Скарга-Бандурова И.С., Щербакова М.Е. Разработка элементов рекомендательной системы на базе веб-аналитики

Статья посвящена анализу методов и инструментов разработки рекомендательной Системи, способной выполнять сбор, анализ, представление и интерпретацию информации о посетителях веб-сайтов с целью улучшения их обслуживания. Рассмотрены сервисы веб-аналитики Google Analytics, их преимущества и недостатки. Предложен вариант собственной системы, построенной с использованием колаборативних фильтрации и для поиска подобных предпочтений пользователей использует косинусные сходство.

Ключевые слова: *веб-аналитика, рекомендательная система, колаборативних фильтрация, онлайн-магазин.*

Kanakin A.E., Skarga-Bandurova I.S., Shcherbakova M.E. Development of elements of a recommender system based on web analytics

The article is devoted to the analysis of methods and tools for developing a recommendatory system capable of collecting, analyzing, presenting and interpreting information about website visitors in order to improve their service. Reviewed by Google Analytics web analytics services, their advantages and disadvantages. A variant of a proprietary system built using colorative filtering and using cosine similarities to search for similar user preferences is proposed.

Keywords: *web analytics, recommender system, colorative filtering, online store.*

Канакін Артур Едуардович, магістрант кафедри комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: agabolik@gmail.com

Скарга-Бандурова Інна Сергіївна, д.т.н., зав. кафедри комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: skarga-bandurova@snu.edu.ua

Щербакова Марина Євгенівна, к.т.н., доц. кафедри комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: m.shcherbakova432@gmail.com

Рецензент: д.т.н., проф. ***Татарченко Г.О.***

Стаття подана 08.09.2018