УДК: 330.4:004.22

В.І. Гриценко, І.М. Онищенко

ЗАСТОСУВАННЯ ІНСТРУМЕНТІВ BIG DATA ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ОНЛАЙН РЕКЛАМИ

Впровадження нових методів та підходів до обробки даних, які отримали назву «Big Data», особливо актуальне для систем із викою завантаженістю. В умовах швидкого потоку даних традиційні пакетні методи моделювання не завжди дають точні та стійкі результати. У даній роботі автором розглянуто онлайновий підхід до прогнозування імовірності кліку користувачем на рекламу та вплив такого підходу на ефективність реклманої кампанії.

***Ключові слова**: інформаційні технології в економіці, економіко-математичне моделювання, алгоритми онлайн навчання, аукціон онлайн реклами, Big Data.*

Внедрение новых методов и подходов к обработке данных, получивших название «Big Data», особенно актуально для систем с высокой загруженностью. В условиях быстрого потока данных традиционные пакетные методы моделирования не всегда дают точные и устойчивые результаты. В данной статье автором рассмотрено онлайновый подход к прогнозированию вероятности клика пользователя по рекламе и влияние такого подхода на эффективность рекламной кампании.

***Ключевые слова**: информационные технологии в экономике, экономико-математическое моделирование, алгоритмы онлайн обучения, аукцион онлан рекламы, Big Data.*

Implementation of new methods and approaches to data processing called “Big Data” is actual especially for high velocity systems. An example of such system is an online

advertising auction, where the number of requests is above 100 per second. In case of high velocity traditional batch learning algorithms not always lead to accurate and stable results. In the article, the author deals with online learning algorithm to predict the Click-Through-Rate for an online ad. After the author compare the result of working of two algorithms and shows the problem of using batch learning.

***Key words**: information technologies in economics, economic-mathematic modeling, online learning algorithms, online ad auction, Big Data*

Актуальність. Сучасний бум інформаційних технологій, особливо рішення пов'язані із обробкою великих обсягів даних, приводять до виникнення швидких та високопродуктивних систем та моделей бізнесу у різних галузях. Однією із таких галузей є реклама в мережі інтернет. Завдяки розвитку методів швидкої обробки даних стало можливим проведення аукціонів онлайн реклами в режимі реального часу.

Але для оцінки вартості реклами сторони аукціону повині оцінювати потенційну вигоду від розміщення свого рекламного блоку у кожному конкретному випадку, а саме – імовірність того, що реклама зацікавить користувача і він перейде за рекламним посиланням.

Слід відзначити, що моделювання процесів із високою швидкістю потоків даних відносить до складу інструментів "Big Data" [1]. Оскільки потрібно спроектувати відповідну інформаційну інфраструктуру та будувати моделі які можуть навчатися на даних великих обсягів та високої частотності.

Аналіз останніх досліджень і публікацій.

Проблема прогнозування Click-Through-Rate або CTR, як показника ефективності онлайн реклами завжди цікавила спеціалістів даної сфери оскільки він впливає на

ціноутворення на ринку реклами в інтернеті, а відповідно і на витрати та прибутки обох сторін.

Спочатку задача прогнозування CTR розв'язувалась на основі кластеризації рекламних банерів відносно їх тексту [2]. Тут автори висувають гіпотезу, що схожі за описанням банери мають подібний CTR.

Пізніше у роботі [3] автори запропонували більш широкий підхід до прогнозування CTR. В якості набору даних для навчання моделі вибираються тільки такі банери, які набрали щонайменше 100 показів за історію, а в якості моделі автори використовують логістичну регресію.

Також слід відмітити наукові дослідження у сфері онлайнових (динамічних) методів машинного навчання, а саме – Ботту Л. та Шалев-Шварц С., а також праці команди авторів із компанії Google під керівництвом МакМахана [4, 5, 6]. Серед відчизняних науковців відзначимо праці Кудінова П.Ю., Полежаєв В.А., Терентьєва А.Н. та Бідюка П.І. [7, 8]

Невирішені проблеми. Недоліки першого підходу – сегментації банерів на групи у тому, що такий метод дає дуже усереднені оцінки. Такі результати можна використовувати як перше наближення, але не кінцевий результат. Суми залишків у даному випадку занадто великі.

Другий підхід не враховує можливості навчання на великих обсягах даних та ігнорує цілий клас банерів із рідкісною аудиторією.

Мета статті. У даній статті основною метою є дослідити інструмент «Big Data», а саме – онлайнові алгоритми машинного навчання, для прогнозування ефективності інтернет реклами в умовах роботи аукціону в реальному часі. Також буде досліджено питання про

традиційні методи машинного навчання, границі їх застосування та порівняння із онлайновими методами.

Постановка завдання. У даній статті поставлено та досліджено такі наукові завдання: дослідити процес роботи аукціону інтернет реклами в реальному часі, побудувати прогноз параметра CTR – як основного показника ефективності інтернет реклами, дослідити можливості онлайнових алгоритмів машинного навчання для оптимізації процесу прогнозування показника CTR.

Виклад основного матеріалу.

Інтернет-реклама як рекламний засіб включає в себе широкий спектр різних інструментів впливу наспоживача. До засобів інтернет-реклами можна віднести веб-сервер, банери, рекламні мережі, електронна пошта, групи новин, пошукові системи і каталоги, інтернет-аукціон, мережі обміну миттєвими повідомленнями, "жовті сторінки". Найкращими можливостями в поданні інформації мають такі засоби інтернет-реклами, як банери, рекламні мережі та електронна пошта.

За типом спрямованості на споживача банерну рекламу можна розділити на кілька класів, а саме: високопродуктивний банер, бренд-банер, банер для просування товару (послуги), банер з високим рівнем відгуку.

Високопродуктивний банер націлений на певну аудиторію. Його завдання - зацікавити і залучити певного відвідувача. Для розробки концепції банера в першу чергу визначаються мета проведеної рекламної кампанії і бажаний кінцевий результат.

Бренд-банер забезпечує запам'ятовування бренду (торгової марки) в певній зоні інтересів. Ефективність бренд-банера не вимірюється кількістю кліків на нього, а безпосередньо залежить від вдалої реалізації та кількості

показів. Основне завдання банера – створити образ компанії чи бренду в пам'яті аудиторії і встановити асоціацію образу з пропонованим товаром або послугою.

Банер для просування товару (послуги) – це банер, ключовим моментом якого є спеціальна пропозиція або знижка. Основне завдання банера - створити стимул до дії (купівлі, відвідування сайту і т. д.). Ефективність банера виражається у відсотку відвідувачів, які вчинили дію.

Банер з високим рівнем відгуку. Основне завдання банера - змусити користувача натиснути на нього.

Ефективність банера виражається в кількості натискань на нього (CTR – click through rate) – відношення користувачів, що перейшли за посиланням, до загального числа показів даного оголошення.

Продаж онлайн реклами може відбуватися кількома способами, один із перших та найбільш розповсюджених – продаж паказів оптом – пакетами по 1000 показів. Такий підхід довів свою простоту та зрозумілість використання для покупців та продавців реклами. Недоліками такого підходу є:

1. неможливість гнучкого таргетування цільової аудиторії, що призводить до неефективних показів;
2. неможливість визначити вартість кожного окремого перегляду реклами;
3. неможливість змінювати рекламний контент у відповідності до характеристи та інтересів користувача;
4. неможливість прогнозувати ефективність рекламного оголошення (CTR).

До вказаного вище необхідно додати інформацію про основні тренди у галузі антернет реклами. Загальносвітова кількість користувачів інтернетом збільшилась із приблизно 100 мільонів у 1996 році до 3 мільярдів у 2014,

при цьому темпи приросту аудиторії не знижуються (Рис.1).

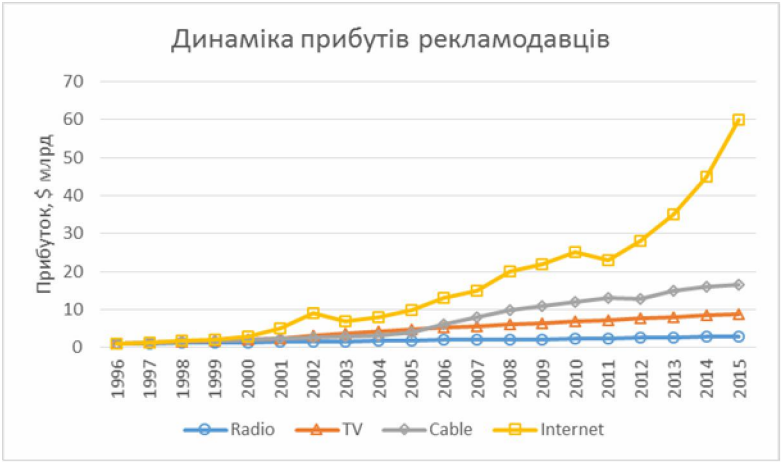
Джерело: [Розроблено автором на основі даних 9]

Рис. 1. Приріст користувачів інтернету та тренд.

При таких темпах зростання аудиторії прямо пропорційно збільшується трафік відвідуваних веб-сторінок, а отже і кількість переглядів онлайн реклами. Зростання аудиторії у свою чергу приводить до зростання прибутку рекламодавців Рис 2.

З іншого боку зростання інтернет аудиторії призводить до збільшення рекламних показів користувачам, що не є цільовою аудиторією замовника реклами та до зниження CTR. Також, очевидною стає сегментація трафіку за якісними характеристиками, основною з яких є кількість переходів за рекламним посиланням – CTR. Власники сайтів хочуть оптимізувати свої прибутки через диференціацію ціни для різних сегментів трафіку – чем вище CTR тим вище ціна за показ реклами.

Таким чином, з'являється необхідність сегментації цільової аудиторії за групами, що неможливо було зробити без застосування надійних методів швидкої обробки великої кількості запитів. Аукціони онлайн реклами починають застосовувати більш технологічні рішення для аналізу даних та сегментації трафіку. З'являється така модель продажу інтернет реклами як аукціони в режимі реального часу. Продаж онлайн-реклами через аукціон в режимі реального часу дозволяє клієнтам купувати вже відому аудиторію та з деякою імовірністю прогнозувати відгук. Після того як користувач відкрив веб-сторінку автоматично починається аукціон. . Компанії змагаються за розміщення реклами, враховуючи місце її появи і те, що їм відомо про потенційного відвідувача сторінки з цифрових слідів, які він залишив у мережі.



Джерело: [Розроблено автором на основі даних 10]

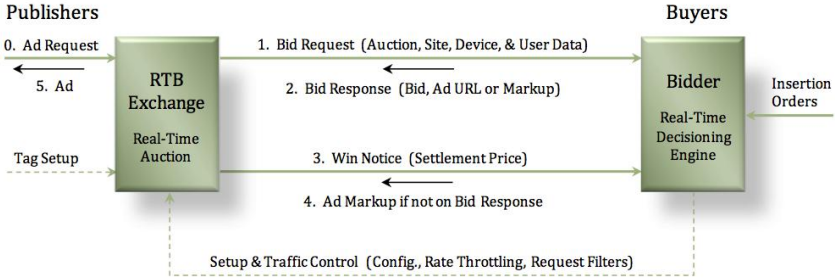
Рис. 2. Динаміка зміни прибутків рекламодавців за сегментами.

Переможець аукціону розміщує рекламу, часто коригуючи її відповідно своїй бізнес-логіці, наприклад у сонячні дні рекламується більше кабріолетів ніж у похмурі. Зазвичай весь процес аукціону та прийняття рішення яку саме рекламу показати займає не більше 150 мілісекунд.

Розглянемо більш детально роботу аукціону з продажу онлайн-реклами в режимі реального часу. Продавцем реклами на аукціоні виступають рекламні площадки, тобто сайти на які заходять користувачі та можна розмістити рекламу. Такі сайти називають publisher. Покупцями реклами є автоматизовані системи купівлі реклами – Demand-Side Platform (DSP). Аукціон проводиться на спеціальних біржах в режимі реального часу – Real-Time Bidding Exchange (RTB).

Аукціон починається як тільки інтернет сторінка із рекламним блоком починає завантажуватись у браузер користувача (Рис. 3):

1. RTB exchange передає потенційним покупцям інформацію про сторінку (URL), рекламний блок (розмір, позиція, вимоги до реклами) та анонімний ідентифікатор інтернет-користувача.
2. Покупці DSP – перевіряють наявну інформацію про користувача (місце знаходження, час, пристрій, вік та ін.) та визначають ціну яку вони згодні заплатити за показ своєї реклами даному користувачу. При цьому покупці можуть запитувати інформацію у спеціальних базах даних – Data Management Platform (DMP).
3. RTB exchange приймає ставки та визначає переможця. Аукціон займає близько 100 мілісекунд.
4. Реклама з максимальною ставкою буде показана на сторінці.



Джерело: [11]

Рис. 3. Схема роботи аукціона інтернет-реклами в режимі рекламного часу.

Враховуючи кількість запитів, які проходять через аукціон – понад 100 запитів за секунду та вказану вище швидкість обробки даних стає очевидним необхідність у надійних та швидких системах оцінки запитів та прийняття рішень на основі отриманих оцінок. Необхідно навчити модель на даних які надходять із великою швидкістю та у великих обсягах.

Далі розглянемо задачу прогнозування ефективності банера, а саме кількості кліків на рекламне оголошення (CTR) та які переваги прононують методи Big Data.

Визначення: Big Data в інформаційних технологіях – серія підходів, інструментів та методів обробки структурованих та неструктурованих даних великих обсягів і різноманітності для отримання результатів, які:

1) легко сприймаються людиною,
2) ефективні в умовах неперервного приросту, розподілення по численним вузлам обчислювальної мережі.

В якості характеристик, які визначають поняття великих даних, відзначають «три V»:

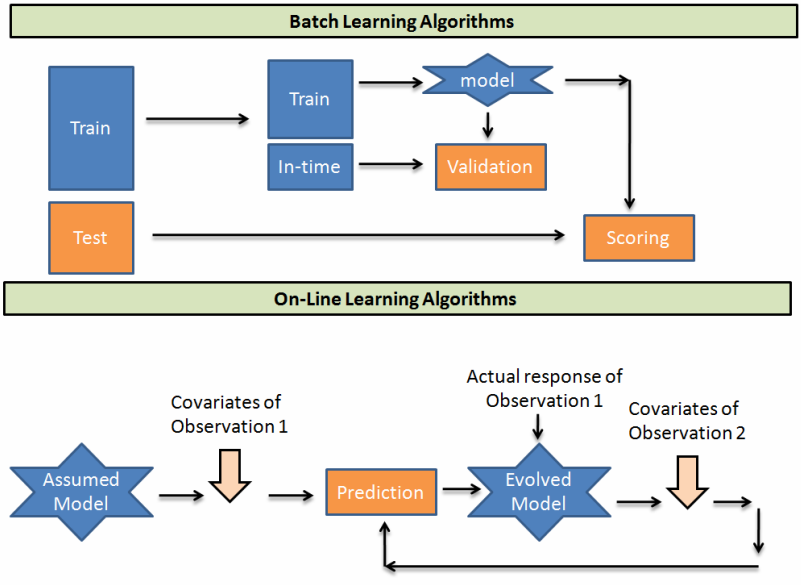
1) Volume – об’єм;

2) Velocity – швидкість, як у розумінні швидкості приросту, так і необхідності швидкої обробки та отримання результату;

3) Variety – різноманітність, у розумінні можливості одночасної обробки різних типів даних. [12]

Традиційно для розв'язання таких задач використовують методи машинного навчання основані на певному фіксованому наборі даних – це так званий пакетний (batch) підхід. При цьому усі дані доступні одразу і можуть бути оброблені на одному обчислювальному вузлі. Також пакетний підхід означає, що модель спочатку була навчена на певному наборі даних – training dataset, а потім тестується на тестовому наборі даних – test dataset та використовується для прогнозування на практиці (Рис. 4). В основі такого підходу лежить гіпотеза про те, що структура даних та статистичні співвідношення між параметрами моделі не змінюються в часі.

Спроби розв'язати задачу прогнозування CTR пакетними методами прогнозування призводили до нестійких у часі результатів, оскільки розмір вибірки для навчання від онлайн аукціону сягає кількох мільйонів записів за 1 день. Зміна структури вибірки досить відчутна при вивченні із розбиттям по дням тижня – тому обмежитись набором даних одного дня неприпустимо. Також постійно присутні сильні впливи, наприклад, публікація резонансної новини на сайті, яка значно мінює тренд переходу за посиланням та саму структуру даних. При збільшенні періоду кількість даних зростає катастрофічно. Але навіть обробивши дані і отримавши більш надійний прогноз, похибка швидко зростає враховуючи динамічність системи.



Джерело: [13]

Рис. 4. Порівняння пакетного та онлайнового методів навчання моделі.

Вказані вище проблеми можна вирішити із застосуванням алгоритмів для моделювання, які навчаються постійно і одночасно дозволяють отримувати прогнозовані значення (Рис. 4). При цьому прогнозуючи величину CTR для нових даних і навчаючись на них, як тільки для них стає відомий факт «відгуку», таким чином постійно коректуючи параметри системи для стабілізації точності прогнозування у часі.

Розглянемо далі застосування алгоритму Follow The Regularized Leader (FTRL) для задачі прогнозування CTR [6]. Даний алгоритм базується на тому, що на кожному кроці вибирається такий набір параметрів який призводить до найменшої похибки на даному кроці:

$$w_t = \arg min \sum_{i=1}^{t-1} v_i(w) + R(w)$$

Функція втрат має вигляд:

$$v_t(w) = \|w - x_t\|^2$$

У випадку лінійної функції оптимізації функція втрат має вигляд:

$$v_t(w) = \langle w, z_t \rangle$$

Оскільки у випадку пронозування події переходу за рекламним посиланням ми маємо справу з бінарною залежною змінною, то зручно використовувати логарифмічну функцію втрат:

$$v_t = (\sigma(w \cdot x_t) - y_t)x_t$$

де σ – сігмоїдальна функція:

$$\sigma(a) = \frac{1}{1 + e^a}$$

Також потрібно вибрати функцію регуляризації. Загальноприйнятими є такі типи регуляризації – L0, L1 та L2. [14]

Нехай ми вибрали наступну функцію регуляризації:

$$R(w) = \frac{1}{2\eta}\|w\|^2$$

для деякого $\eta > 0$. Тоді ітерація алгоритму навчання матиме вигляд:

$$w_{t+1} = -\eta \sum_{i=1}^{t} z_i = w_t - \eta z_t$$

Останню рівність можна також переписати у вигляді:

$$w_{t+1} = w_t - \eta \nabla v_t(w_t)$$

що відповідає рівнянню алгоритму покрокового градієнтного спуску.

Остаточно формула алгоритму з регуляризацією виглядає так:

$$w_{t,i} = \begin{cases} 0, & \text{при } |z_i| \le \lambda_1 \\ -\left(\frac{\beta + \sqrt{n_i}}{\alpha} + \lambda_2\right)^{-1} (z_i - sign(z_i)\lambda_1), & \text{при } |z_i| > \lambda_1 \end{cases}$$

де α, β – вхідні параметри моделі, що відповідають за швидкість навчання, λ_1, λ_2 – параметри моделі, що відповідають за силу регуляризації L1 та L2 відповідно. Параметри z, n обраховуються на кожному кроці ітерації разом із коефіцієнтами моделі:

$$v_i = (p_t - y_t)x_i$$
$$\sigma_i = \frac{1}{\alpha}\left(\sqrt{n_i + v_i^2} - \sqrt{n_i}\right)$$
$$z_i = z_i + v_i - \sigma_i w_{t,i}$$
$$n_i = n_i + v_i^2$$

На відміну від алгоритму покрокового градієнтного спуску, який під час своєї роботи повинен зберігати коефіцієнти моделі w, FTRL алгоритм зберігає параметр z а потім розраховує коефіцієнти із використанням коефіцієнтів швидкості навчання та регуляризації. Якщо покласти $\eta = const, \lambda_1 = 0$ – то ми отримаємо алгоритм градієнтного спуску.

Модель FTRL відноситься до класу жадібних алгоритмів (greedy algorithm). Даний клас алгоритмів базується на прийнятті локально оптимального рішення на кожному кроці з метою прийти до глобального оптимуму в кінці. [15]

Для прогнозування використовувався розглянутий вище алгоритм Follow The Regularized Leader з по координатним спуском та L1, L2 – регуляризацією. Для порівняння використовувалась модель для пакетного навчання, максимально схожа на розглянуту вище лінійну модель FTRL – логістична регресія. Модель навчалась на

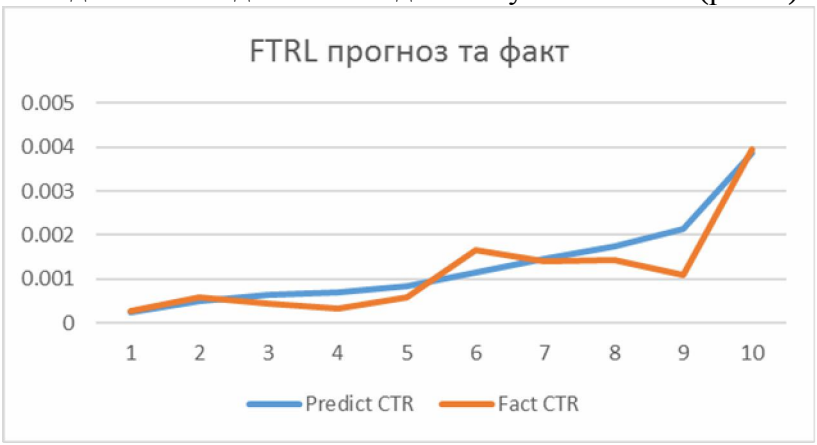
10 мільйонах спостережень, а тестувалась на 3 мільйонах. Далі наведена порвняльна таблиця результатів роботи моделей:

Таблиця 1

Порівняння результатів роботи пакетного та онлайногового алгоритмів.

	Логітична модель	**FTRL модель**
Час навчання	5 годин 46 хвилин	15 хвилин
Джині	62%	65%
MAE (середня абсолютна похибка)	0,004	0,003

Оскільки результати прогнозу логістичної моделі та FTRL моделі є приблизно рівними то наведемо результати для FTRL моделі. Графік порівняння прогнозу та факту CTR для обох моделей виглядає наступним чином (рис. 5):



Джерело: Розроблено автором

Рис.5. Прогнозні та фактичні значення імовірностей CTR.

Показники точності моделювання приблизно однакові для двох підходів, але час роботи візрізняється критично. Якщо до цього додати той факт що в онлайн аукціоні 10 мільйонів записів може зібратися за кілька годин – то використання пакетного підходу для постійної підтримки адекватності результатів стає неприйнятним.

Завдяки онлайновому методу навчання модель коректується з постійним кроком. Тобто інформація, що приходить з новими спостереженнями впливає на параметри моделі так само як і попередні дані. Таким чином параметри моделі постійно корегуються в режимі реального часу і постійно залишаються актуальними, що позитивно відзначається на точності прогнозу.

На основі прогнозів CTR можуть формуватися стратегії boosting – покращення рекламних кампаній на основі відбору трафіку з вищім рівнем клікабельності при чому точку відсікання для прийняття рішення доцільно зробити плаваючою. За такого підходу онлайн-аукціон збільшує свої прибутки від продажу кліків, компанії отримують більше потенційних клієнтів і підвищують продажі або впізнавання свого бренду.

Висновки. Пакетне навчання показує гарні результати та пропонує велику кількість різних моделей, але у випадку коли дані змінюються із великою швидкістю та у великих обсягах пакетне навчання може не впоратися із завданням. Проблему навчання моделі на Big Data наборах вирішують за допомогою онлайн алгоритмів. Онлайн алгоритми дозволяють створювати динамічні системи, які працюють в режимі реального часу, постійно коригують параметри моделі у ситуації, коли дані надходять із великою швидкістю.

Необхідно також відзначити, що розвиток онлайн алгоритмів зробив великий вклад у розвиток технологій

Big Data значно розширивши наявні інструменти для розв'язання практичних задач. Як інструмент Big Data онлайн алгоритми незамінні при побудові інтелектуальних систем які працюють у режимі реального часу, як наприклад, аукціони онлайн реклами чи рекомендаційні системи

Список використаних джерел

1. Майер-Шенбергер Виктор. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим/ Виктор Майер-Шенбергер, Кеннет Кукьер; пер. с англ. Инны Гайдюк. – М.: Манн, Иванови Фербер, 2014. – 240 с.
2. M. Regelson and D. Fain. Predicting click-through rate using keyword clusters. In Proceedings of the Second Workshop on Sponsored Search Auctions, volume 9623. Citeseer, 2006.
3. M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In Proceedings of the 16th international conference on World Wide Web, pages 521–530. A CM, 2007.
4. Shalev-Shwartz, Shai. "Online Learning and Online Convex Optimization". Foundations and Trends in Machine Learning. 2011. pp. 107–194.
5. Gilles Gasso. Batch and online learning algorithms for nonconvex Neyman-Pearson classification / Gilles Gasso, Aristidis Pappaioannou, Marina Spivak, Leon Bottou / ACM Transaction on Intelligent System and Technologies, 2(3), 2011.
6. H Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. InInternational Conference on Artificial Intelligence and Statistics, pages 525–533, 2011.
7. Кудинов П. Ю. Динамическое обучение распознаванию статистических таблиц / Кудинов П. Ю., Полежаев В.А. / Доклады 8-й Международной конференции «Интеллектуализация обработки информации» ИОИ-2010 (Республика Кипр, г. Пафос, 17-24 октября 2010). – М.: МАКС Пресс, 2010. – С. 512-515.
8. Бидюк П.И. Построение и методы обучения Байесовских сетей /Бидюк П.И., Терентьев А.Н., Гасанов А.С./ Кибернетика и системный анализ, – 2005. – № 4. – С. 133 – 147.

9. Internet live stats. [Електронний ресурс] – режим доступу: http://www.internetlivestats.com/internet-users/
10. Douglas Galbi (purple motes) and IAB Internet Advertising Revenue Report, FY 2015. [Електронний ресурс] – режим доступу: https://www.scribd.com/document/310075259/IAB-Internet-Advertising-Revenue-Report
11. Что такое Real-Time Bidding. [Електронний ресурс] – режим доступу: http://konverta.ru/how
12. Канаракус, Крис. Машина Больших Данных. Сети, № 04, 2011. [Електрорнний ресурс] – режим доступу: http://www.webcitation.org/6AOq8Azix
13. Introduction to online machine learning: Simplified. [Електронний ресурс] – режим доступу: http://www.analyticsvidhya.com/blog/2015/01/introduction-online-machine-learning-simplified-2/
14. Riedman J. H. Regularization paths for generalized linear models via coordinate descent / Riedman J. H., Hastie T., Tibshirani R. / Journal of Statistical Software. 2010. Vol. 33, no. 1. pp. 1–22
15. Кормен Т. Алгоритмы: построение и анализ / Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К./ 2-е издание, М.:Вильямс, 2005, с. 442-478.

УДК 330.4 **О.І. Ляшенко, К.І. Крицун**

ДОСЛІДЖЕННЯ ДИНАМІКИ ФОНДОВОГО ІНДЕКСУ ПФТС НА ФІНАНСОВОМУ РИНКУ УКРАЇНИ НА РІЗНИХ ЧАСОВИХ ВІКНАХ З 2001 ПО 2016 РОКИ

У статті досліджено динаміку фондового індексу ПФТС з 2001 по 2016 роки. Застосовано мультифрактальний аналіз та R/S-аналіз, як інструменти аналізу динаміки фінансових часових рядів. Здійснено розрахунок індексу Херста за допомогою програмного пакету Gretl. Виявлено мультифрактальні властивості ряду за допомогою програми SpectrAnalyzer. Графічно