УДК 519.6

Аспірант Павлов Д.Г.

**Національний технічний університет України
«Київський політехнічний інститут»**

ЗАСТОСУВАННЯ АЛГОРИТМУ "ГУСЕНИЦЯ"-SSA ДЛЯ БОРОТЬБИ ІЗ МЕРЕЖЕВИМ ШАХРАЙСТВОМ В СИСТЕМІ КОНТЕКСТНОЇ РЕКЛАМИ

Abstract

Dmytro Pavlov, graduate student
"Caterpillar"-SSA for fraud detection in contextual advertising system

This paper concerns the task of fraud detection in contextual advertising system. Author proposes using "Caterpillar"-SSA algorithm for click (impression) time series decomposition. By comparing obtained trend components, it is possible to identify some fraud attacks.

Вступ

Все більше рекламодавців віддають перевагу рекламі в мережі Інтернет, зокрема, контекстній рекламі. Контекстна рекламна кампанія в мережі є дешевою в порівнянні із рекламою в традиційних засобах масової інформації, а її безпосередня направленість на цільову аудиторію потенційних клієнтів забезпечує швидку віддачу рекламних фондів. Ці фактори підвищують популярність мережевої реклами серед рекламодавців представників середнього та, особливо, малого бізнесу.

Однак реклама в Інтернеті є вразливою до специфічного різновиду мережевого шахрайства. В залежності від моделі проведення рекламної кампанії: cost per click – рекламодавець сплачує кожен перехід по рекламному оголошенню або cost per mille – рекламодавець платить за кожну тисячу показів оголошення, шахраї можуть генерувати штучні кліки (склікування) або покази (споказування) із метою витрачення рекламного бюджету рекламодавця.

Рекламні мережі розробляють спеціальні системи захисту, активно використовується фільтрація потоку кліків (показів) в режимі реального часу [1]. Однак, рівень мережевого шахрайства залишається досить високим. Про це, приміром, свідчать різноманітні звіти по рівню недійсних

кліків в мережі [2] та скарги на якість захисних систем пошукових мереж [3].

Тому розроблення нових підходів для аналізу потоку кліків (показів) з метою визначення наявності шахрайства є актуальною задачею. В роботі [4] було запропоновано використовувати вейвлет-перетворення для фільтрації мережевого трафіку. Головною особливістю вейвлет-аналізу є те, що він дозволяє виділяти різкі сплески в сигналах, а, отже, дозволяє визначити початок склікування чи споказування, як різку зміну в загальному потоці.

В даній роботі автор пропонує використовувати метод "Гусениця"-SSA для аналізу часових рядів кліків та показів. Цей метод дозволяє ефективно розкладувати сигнал на трендову, сезонну та шумову компоненти.

Якщо при порівнянні трендових компонент двох або більше рекламних кампаній, що використовують однакові або пов’язані між собою ключові слова, визначається наявність розбіжностей, це може свідчити про наявність шахрайських дій в мережі.

Постановка задачі

Метою даної роботи є дослідження потенціальної можливості використання методу "Гусениця"-SSA для визначення наявності мережевого шахрайства в системі контекстній реклами.

Короткий опис базового алгоритму "Гусениця"-SSA

Базовий алгоритм "Гусениця"-SSA складається з двох етапів: розкладення та відновлення [5]. Під час першого етапу для початкового часового ряду $F = f_0 \ \dots \ f_{N-1}$ будується траєкторна матриця, стовпчики якої мають вигляд $X_i = f_{i-1} \ \dots \ f_{i+L-2}^T$, $1 \le i \le N - L + 1$. L – довжина вікна, єдиний параметр даного методу.

Наступним кроком проводиться сингулярний розклад траєкторної матриці, в результаті якого вона представляється у вигляді суми елементарних матриць одиничного рангу:

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d . \qquad (1)$$

Після цього проводиться групування елементарних доданків в сумі (1). Таким чином, множина індексів $1,...,d$ розділяється на m

підмножин, що не перетинаються $I_1,\ldots,I_m$, а траєкторна матриця представляється у вигляді:

$$\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}. \quad (2)$$

Далі за допомогою процедури діагонального усереднення кожен із доданків-матриць суми (2) переводиться в ряд $\tilde{F}_j$, а отже початковий часовий ряд теж розкладається в суму $F = \tilde{F}_{I_1} + \ldots + \tilde{F}_{I_m}$.

За рахунок правильного підбору довжини вікна та успішного групування доданків із суми (1) можна виділити із початкового сигналу трендову, коливальні та шумову складові.

Оскільки довжина вікна L є єдиним параметром методу, її визначення є одним з найголовніших кроків алгоритму. Для досягнення найкращого результату бажано обирати великі значення L, проте вони не повинні перевищувати половини довжини ряду. Для успішного відділення коливальної компоненти від інших також необхідно, щоб довжина вікна була кратною її періоду.

Застосування алгоритму "Гусениця"-SSA для аналізу часових рядів кількості кліків

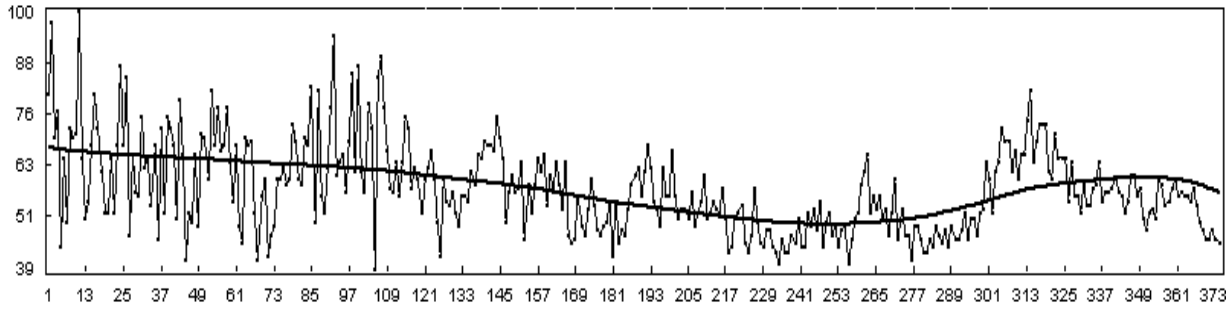
На рис. 1 наведено приклад аналізу трафіку за двома ключовими словами "комп’ютер" та "монітор" методом "Гусениця"-SSA. Початкові часові ряди представляють собою тижневу кількість кліків по оголошенню та мають довжину 353 відліки. Оскільки даним часовим рядам можуть бути властиві сезонні коливання, в якості довжини вікна було обрано число 156, що дозволяє відділяти як річні та піврічні коливання (52 та 26 тижнів) так і коливання, що властиві порам року (12 тижнів). В якості трендових складових було обрано перші 2 доданки із суми (1).

На рис. 1 видно, що в кінці спостережень трендова складова першого ряду почала зростати, на відміну від трендової складової другого. Це може свідчити про наявність склікування в зазначений період.

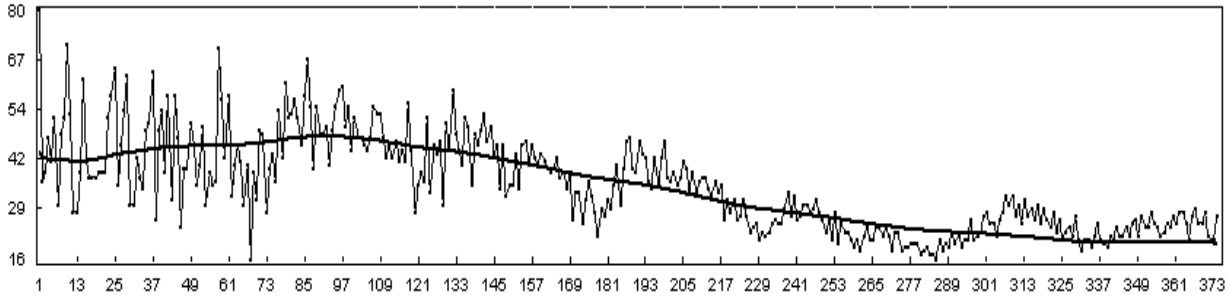
Висновки

Проведене дослідження показало, що метод "Гусениця"-SSA може використовуватись для виділення складових часових рядів кількості кліків або показів. Порівнюючи отримані трендові складові часових рядів декількох пов’язаних між собою причинно-наслідковим зв’язком

рекламних кампаній, можна зробити висновки щодо наявності шахрайства в системі контекстної реклами.



а)



б)

Рис. 1. Часові ряди кількості кліків в тиждень за словами "комп'ютер" (а) і "монітор" (б) та їх трендові компоненти

Література

1. How does Google detects invalid clicks? [Electronic resource]. – Режим доступу: http://adwords.google.com/support/aw/bin/answer.py?hl=en&answer=61144
2. *Kitts B., LeBlanc B., Meech R., Laxminarayan P.* Click-fraud. [Electronic resource]. – Режим доступу: http://www.asis.org/Bulletin/Dec-05/clickfraud.html.
3. *Сазанов В.М.* Виртуальная школа компьютерных технологий. Лекция 15 [Electronic resource]. – Режим доступу: http://v-school.narod.ru/INI/ini.htm
4. *Chertov O., Malchykov V., Pavlov D.* Non-dyadic wavelets for detection of some click-fraud attacks // 2010 International Conference on Signals and Electronic Systems (ICSES). – 2010. – pp. 401 – 404.
5. *Голяндина Н.Э.* Метод "Гусеница"-SSA: анализ временных рядов. Учеб. пособие. – СПб., 2004. – 76 с.