Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/
Том 7, №3 (2015) http://naukovedenie.ru/index.php?p=vol7-3
URL статьи: http://naukovedenie.ru/PDF/41TVN315.pdf
DOI: 10.15862/41TVN315 (http://dx.doi.org/10.15862/41TVN315)

УДК 004.021

Савин Андрей Сергеевич
ООО «Махуру»
Российская Федерация, Москва[1]
Программист
E-mail: asavin@mahuru.ru

Хохлов Алексей Анатольевич
ФГБОУ «Российский университет дружбы народов (РУДН)»
Российская Федерация, Москва
Доцент
Кандидат физико-математических наук
E-mail: khokhlov_aa@pfur.ru

Четов Артур Игоревич
ФГБОУ «Российский университет дружбы народов (РУДН)»
Российская Федерация, Москва
Студент
E-mail: karelia_90_@mail.ru

Анализ временных рядов в приложении к изучению поведения покупателей

[1] 115419, Москва, Орджоникидзе, 3, к. 118

Аннотация. В настоящей работе рассматривается задача применения алгоритма анализа временных рядов «Гусеница» к исследованию поведения клиентов магазина. Авторами была проделана работа по созданию и внедрению программного комплекса, состоящего из мобильного приложения для клиентов, позволяющего собирать статистику поведения клиентов, серверной части для хранения и обработки данных, а также работа по анализу полученной за пять месяцев работы статистики с применением методов анализа временных рядов. Были выяснены закономерности, что накопление большего количества данных может позволить формировать эффективные стратегии взаимодействия с клиентами. Результаты анализа и выводы представлены в работе.

Ключевые слова: SSA; метод «Гусеница»; анализ поведения клиентов; временной ряд.

Ссылка для цитирования этой статьи:

Савин А.С., Хохлов А.А., Четов А.И. Анализ временных рядов в приложении к изучению поведения покупателей // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 7, №3 (2015) http://naukovedenie.ru/PDF/41TVN315.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ. DOI: 10.15862/41TVN315

Развитие мобильного интернета, повсеместное распространение смартфонов, удешевление связи и другие факторы предоставили новые возможности для бизнеса по взаимодействию со своими клиентами. Теперь, чтобы донести до клиента информацию, предоставить ему скидку, дать возможность зафиксировать покупку или факт использования услуги, достаточно разработать и опубликовать мобильное приложение, которое может установить любой человек. Помимо выполнения своей прямой цели – коммуникация с клиентами в обе стороны при помощи различных технологий, таких как PUSH сообщения, такие инструменты позволяют собирать и анализировать подробную статистику о поведении своих клиентов.

Например, можно собирать статистику о том, когда, в какое время, в какую погоду, при каком курсе валют было сделано то или иное количество покупок, какой возрастной категорией, новые это клиенты или старые – факторов для анализа может быть много.

Такого рода задачи являются актуальными на данный момент, так как в результате можно формировать эффективные стратегии коммуникации с клиентами, предлагать адресные предложения с высокой конверсией, избавить клиентов от ненужной им назойливой рекламы – одним словом, предлагать клиентам именно то, что им нужно (на основании статистического портрета клиента) именно в данный момент.

На сегодняшний день системы, которые позволяют решать такие задачи, используются только в крупных компаниях и являются очень дорогими. Авторы поставили перед собой задачу создать программный комплекс, который мог бы интегрироваться с большинством распространенных кассовых систем, был сравнительно недорогим и позволял, используя методы и алгоритмы, относящиеся к концепции Big Data [1, 2], проводить анализ данных, прогнозирование тех или иных процессов и формировать рекомендации для бизнеса [3].

В данной статье описан проведенный с декабря 2014 года по май 2015 года эксперимент по разработке и внедрению прототипа такого комплекса в крупный сетевой магазин одежды, который включал в себя разработку мобильного приложения для платформ iOS/Android, разработку базы данных и системы управления данными, обработку полученных данных при помощи алгоритма анализа временных рядов «Гусеница», известный также, как SSA (Singular Spectrum Analysis) [4, 5, 6].

Мобильное приложение, которое устанавливали клиенты магазина, выполняло различные функции. Для анализа в данной работе важны только некоторые из них – получение некоторого бонуса за регистрацию и фиксация покупок. Данные в интерактивном режиме передаются на сервер, где хранятся в специально спроектированной базе данных, позволяющей делать в любой момент времени удобные выборки.

После накопления определенной статистики (около 30 000 установок) авторами был проведен анализ полученных данных, представленных в виде временных рядов. Для этого был выбран алгоритм SSA.

Алгоритм SSA не требует присутствия особенных характеристик у исследуемого временного ряда, будь то стационарность, знания модели, наличия периодических составляющих и других. При этом SSA успешно решает такие задачи, как, выделение трендов, обнаружение периодик, сглаживание ряда, построение полного разложения ряда в сумму тренда, периодик и шума и задачи фильтрации, поэтому он был использован авторами – априорной информации о данных не было.

Алгоритм стандартного метода SSA хорошо известен и изучен [4], поэтому опишем его вкратце. Из исходного одномерного временного ряда строится траекторная матрица, размерность которой определяется параметром, зависящим от условий конкретной задачи –

длина гусеницы. Небольшая длина гусеницы позволяет учесть меньше информации о ряде, большая длина гусеницы требует больших вычислительных ресурсов. Столбцами траекторной матрицы являются скользящие отрезки длиной, равной длине гусеницы. После некоторых преобразований, опционально включающих в себя процедуры нормирования и центрирования, строится квадратная матрица, содержащая в себе информацию об исходном временном ряде. Далее производится сингулярное разложение этой матрицы на сумму элементарных матриц, каждая из которых задается набором из собственного числа и двух сингулярных векторов — собственного и факторного. Таким образом, исходный временной ряд разлагается на интерпретируемы аддитивные составляющие. В зависимости от условий задачи производится отбор главных компонент, по которым при помощи процедуры ганкелизации восстанавливается временной ряд. Непосредственно алгоритм описан ниже.

Рассмотрим временной ряд $\left\{x_i\right\}_{i=1}^{N}$, образованный последовательностью N равноотстоящих значений некоторой функции f_t:

$$x_i = f\left[i\right] = f((i-1)\Delta t), \tag{0.1}$$

где $i = 1, 2, ..., N$.

В качестве примера такой функции можно привести, например, курсы американского доллара, отмечаемые каждый час в течение года. Тогда $N = 365*24 = 8760$.

Задача работы заключается в анализе временного ряда (разложении на главные компоненты, их отбор, восстановление и последующий анализ).

Сначала производится преобразование одномерного ряда в многомерной. Выберем некоторое число $M < N$, называемое длиной гусеницы, и представим первые M значений последовательности f в качестве первой строки матрицы X. В качестве второй строки матрицы берем значения последовательности с x_2 по x_{M+1}. Последней строкой с номером $k = N - M + 1$ будут последние M элементов последовательности : $x_k, x_{k+1}, ..., x_N$:

$$X = (x_{ij})_{i,j=1}^{k,M} = \begin{pmatrix} x_1 & x_2 & \cdots & x_M \\ x_2 & x_3 & & x_{M+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_k & x_{k+1} & \cdots & x_N \end{pmatrix} \tag{0.2}$$

Далее вычисляются средние арифметические значения и стандартные отклонения по столбцам матрицы X

$$\overline{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{i+j-1} \tag{0.3}$$

$$s_j = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_{i+J-1} - \overline{x}_j)^2} \tag{0.4}$$

Обозначим через $X^* = (x_{ij}^*)_{i,j=1}^{k,M}$ матрицу, полученную из X в результате центрирования по столбцам и нормирования на стандарты s_j:

$$x_{ij}^* = (x_{ij} - \overline{x}_j)/s_j;\ i = 1, ..., k;\ j = 1, ..., M \tag{0.5}$$

Операции центрирования и нормирования не являются обязательными. Далее вычисляется матрица

$$R = (1/k)(X^*)^T X^* \tag{0.6}$$

Следующий шаг состоит в вычислении собственных чисел и собственных векторов матрицы R, т.е. разложение ее

$$R = P\Lambda P^T, \tag{0.7}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_M \end{pmatrix} \tag{0.8}$$

- диагональная матрица собственных чисел и

$$P = (p_1, p_2, ..., p_M) = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{M1} \\ p_{12} & p_{22} & \cdots & p_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1M} & p_{2M} & \cdots & p_{MM} \end{pmatrix} \tag{0.9}$$

- ортогональная матрица собственных векторов матрицы R. При этом должны выполняться следующие соотношения:

$$\begin{gathered} P^T = P^{-1}; P^T P = PP^T = I_M, \\ \Lambda = P^T RP, \sum_{i=1}^{M} \lambda_i = M, \\ \prod_{i=1}^{M} \lambda_i = \det R \end{gathered} \tag{0.10}$$

Матрицы Λ и P совместно имеют множество интерпретаций, основанных на анализе главных компонент (АГК) [7]. В частности, матрицу P можно рассматривать как матрицу перехода к главным компонентам:

$$X^* P = Y = (y_1, y_2, ..., y_M). \tag{0.11}$$

Далее необходимо упорядочить матрицу собственных значений по возрастанию и пересортировать соответственно матрицу собственных векторов. Каждое собственное значение (ГК) вносит свой «вклад» в исследуемый процесс, и для проведения анализа пользователь должен иметь возможность выбрать для дальнейшей работы некоторые из них, в зависимости от условий задачи. Таким образом, после этого этапа остается $r < M$ собственных значений и соответствующих им собственных векторов.

Следующим ключевым элемента метода «Гусеница» является процедура восстановления. Эта процедура основана на следующих достаточно простых соотношениях.

Из ортогональности матрицы P следует, что при умножении матрицы главных компонент Y на P^T восстанавливается матрица X^*, при этом получается разложение

$$X^* = YP^T = (y_1, \ldots y_M)\begin{pmatrix} p_1^T \\ p_2^T \\ \vdots \\ . \\ p_M^T \end{pmatrix} = \sum_{l=1}^{M} y_l p_l^T = \sum_{l=1}^{M} X_l^* \qquad (0.12)$$

нормированной и центрированной матрицы X^* в сумму матриц X_l^*, каждая из которых порождена одним собственным вектором матрицы R. Далее производится денормировка X^* с помощью умножения этой матрицы на диагональную матрицу S, состоящую из выборочных средних:

$$X = \bar{x}1_k^T + X^* S = X_0 * + \sum_{l=1}^{M} X_l{}^* S = \sum_{l=0}^{M} X_l{}^* S \qquad (0.13)$$

В результате получается исходная матрица диагональной структуры в виде суммы $M+1$ матриц. Переход к исходному ряду формально может быть осуществлен усреднением по побочным диагоналям. Обозначим через A этот оператор усреднения

$$x = A(X) = \sum_{l=0}^{M} A(X_l{}^* S) \qquad (0.14)$$

Ранее авторами была разработана эффективная реализация алгоритма SSA, которая использовалась при обработке данных [8].

Анализировались такие данные, как количество покупок по дням, количество потраченных денег и для примера параллельно анализировался курс доллара в эти дни. Анализ позволил выявить определенные закономерности. Например, количество денег, которые пользователи мобильных приложений тратили в магазине, имело зависимость от курса доллара, причем обратную. Чем меньше был курс доллара, тем больше денег тратили покупатели. При этом надо отметить, что цены в магазинах от курса доллара не зависели – товар был закуплен осенью по фиксированному курсу.

При этом интересная особенность заключается в том, что количество потраченных денег не зависит прямо от количества покупок (во всяком случае, такой зависимости проследить не удалось, проанализировав поведение восстановленного ряда с разными параметрами SSA – длиной гусеницы и различными отобранными ГК). Таким образом, можно предположить, что, когда клиенты приходят в магазин при низком курсе доллара, они покупают больше дешевых вещей, а когда при высоком – это более осмысленные покупки и люди покупают более дорогие вещи.

Представленные результаты показывают, что задача исследования поведения покупателей при помощи анализа временных рядов даже в самом просто виде позволяет предлагать бизнесу маркетинговые рекомендации – например, в случае высокой волатильности иностранной валюты, на ее ослаблении можно выкладывать в продажу и делать акции на более дешевый товар, который люди будут покупать активнее, а при повышении курса валюты скидки на эти товары можно убирать, так как люди приходят в магазины за более дорогими товарами.

В дальнейшем авторы планируют проводить работу по накоплению статистики за счет подключения к системе новых магазинов, а также учитывать при анализе такие данные, как возраст и пол клиентов, структуру чека, время суток и другие параметры. Для анализа планируется использование методов CSSA и MSSA [9, 10].

ЛИТЕРАТУРА

1. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим; пер. с англ. Инны Гайдюк. — М.: Манн, Иванов и Фербер, 2014. – 240с.
2. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity // McKinsey Global Institute. 2011, 143 p.
3. Петров В.А., Савин А.С., Хохлов А.А., Четов А.И. Анализ временных рядов методом «Гусеница»-SSA в Big Data. Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: материалы Всероссийской конференции с международным участием. Москва, РУДН, 20–24 апреля 2015 г. — Москва: РУДН, 2015. — 332 с.: ил, с. 305-307.
4. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques, CHAPMAN & HALL/CRC, 2001.
5. Golyandina N., Zhigljavsky A. Singular Spectrum Analysis for Time Series, Berlin: Springer, 2013. — 120 p.
6. Голяндина Н.Э. Метод «Гусеница»-SSA: анализ временных рядов: Учеб. пособие. СПб: Изд-во СПбГУ, 2004. 76 с.
7. Данилов Д.Л. Главные компоненты временных рядов: метод «Гусеница» / Под ред. Д.Л. Данилова, А.А. Жиглявского. – СПб: Пресском, 1997. — 308 с.
8. Савин А.С., Хохлов А.А. Оптимизация алгоритма Singular Spectrum Analysis для ARM процессоров мобильных устройств // Интернет-журнал «Науковедение», 2014 №2(21) [Электронный ресурс]-М.: Науковедение, 2014. – Режим доступа: http://naukovedenie.ru/PDF/110TVN214.pdf, свободный. – Загл. с экрана. - Яз. рус., англ.
9. Петров В.А., Савин А.С., Хохлов А.А., Четов А.И. Задача формирования маркетинговых стратегий для ресторанного бизнеса. Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: материалы Всероссийской конференции с международным участием. Москва, РУДН, 20–24 апреля 2015 г. — Москва: РУДН, 2015. — 332 с.: ил, с. 308-309.
10. Голяндина Н.Э., Некруткин В.В., Степанов Д.В. Варианты метода «Гусеница»-SSA для анализа многомерных временных рядов. Труды II Международной конференции «Идентификация систем и задачи управления» SICPRO'03. Москва, 2003, с. 2139-2168.

Рецензент: Ловецкий Константин Петрович, кандидат физико-математических наук, доцент кафедры прикладной информатики и теории вероятностей РУДН.

Savin Andrey Sergeevich
«Mahuru», Ltd
Russian Federation, Moscow
E-mail: asavin@mahuru.ru

Khokhlov Aleksey Anatol'evich
Peoples' Friendship University of Russia (PFUR)
Russian Federation, Moscow
E-mail: khokhlov_aa@pfur.ru

Chetov Artur Igorevich
Peoples' Friendship University of Russia (PFUR)
Russian Federation, Moscow
E-mail: karelia_90_@mail.ru

Time series analysis applied to the study of consumer behavior

Abstract. In this paper we consider the problem of applying the algorithm of time series analysis "Caterpillar" to study the behavior of customers at the shop. The authors developed and implemented software consisting of a mobile application for customers which allows collecting statistics of customer behavior and the server side to store and processing data. Authors analyzed results obtained during five months of work with the use of statistical methods for analyzing time series. Authors clarified dependencies which can allow forming effective strategies for interacting with customers in future. The analysis results and conclusions are presented in the work.

Keywords: SSA; Caterpillar; behavioral analysis; time series.

REFERENCES

1. Mayer-Shenberger V., Kuk'er K. Bol'shie dannye. Revolyutsiya, kotoraya izmenit to, kak my zhivem, rabotaem i myslim; per. s angl. Inny Gaydyuk. — M.: Mann, Ivanov i Ferber, 2014. – 240s.
2. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity // McKinsey Global Institute. 2011, 143 p.
3. Petrov V.A., Savin A.S., Khokhlov A.A., Chetov A.I. Analiz vremennykh ryadov metodom «Gusenitsa»-SSA v Big Data. Informatsionno-telekommunikatsionnye tekhnologii i matematicheskoe modelirovanie vysokotekhnologichnykh sistem: materialy Vserossiyskoy konferentsii s mezhdunarodnym uchastiem. Moskva, RUDN, 20–24 aprelya 2015 g. — Moskva: RUDN, 2015. — 332 s.: il, s. 305-307.
4. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques, CHAPMAN & HALL/CRC, 2001.
5. Golyandina N., Zhigljavsky A. Singular Spectrum Analysis for Time Series, Berlin: Springer, 2013. — 120 p.
6. Golyandina N.E. Metod «Gusenitsa»-SSA: analiz vremennykh ryadov: Ucheb. posobie. SPb: Izd-vo SPbGU, 2004. 76 s.
7. Danilov D.L. Glavnye komponenty vremennykh ryadov: metod «Gusenitsa» / Pod red. D.L. Danilova, A.A. Zhiglyavskogo. – SPb: Presskom, 1997. — 308 s.
8. Savin A.S., Khokhlov A.A. Optimizatsiya algoritma Singular Spectrum Analysis dlya ARM protsessorov mobil'nykh ustroystv // Internet-zhurnal «Naukovedenie», 2014 №2(21) [Elektronnyy resurs]-M.: Naukovedenie, 2014. – Rezhim dostupa: http://naukovedenie.ru/PDF/110TVN214.pdf, svobodnyy. – Zagl. s ekrana. - Yaz. rus., angl.
9. Petrov V.A., Savin A.S., Khokhlov A.A., Chetov A.I. Zadacha formirovaniya marketingovykh strategiy dlya restorannogo biznesa. Informatsionno-telekommunikatsionnye tekhnologii i matematicheskoe modelirovanie vysokotekhnologichnykh sistem: materialy Vserossiyskoy konferentsii s mezhdunarodnym uchastiem. Moskva, RUDN, 20–24 aprelya 2015 g. — Moskva: RUDN, 2015. — 332 s.: il, s. 308-309.
10. Golyandina N.E., Nekrutkin V.V., Stepanov D.V. Varianty metoda «Gusenitsa»-SSA dlya analiza mnogomernykh vremennykh ryadov. Trudy II Mezhdunarodnoy konferentsii «Identifikatsiya sistem i zadachi upravleniya» SICPRO'03. Moskva, 2003, c. 2139-2168.