УДК 004.622

АНАЛИЗ И ОБРАБОТКА ДАННЫХ КОМПАНИИ КАК СПОСОБ ПОВЫШЕНИЯ ЕЕ ПРИБЫЛЬНОСТИ

ANALYSIS AND PROCESSING OF COMPANY DATA AS A METHOD OF PROFITABILITY IMPROVING

Солодкая Н.А.
кандидат технических наук, доцент,
доцент кафедры информационных систем,
Украинский государственный химико-технологический университет

Шаповалов Д.Д.
студент,
Украинский государственный химико-технологический университет

Solodka Nataliia
Candidate of Technical Sciences, Associated Professor,
Senior Lecturer at the Department of Information Systems,
Ukrainian State University of Chemical Technology

Shapovalov Danylo
Student,
Ukrainian State University of Chemical Technology

В статье рассмотрен пример ведения и анализа статистики продаж компании с целью формирования на его основе и реализации алгоритмов подготовки данных для эффективной работы моделей продаж. Выделены факторы, оказывающие влияние на изменение объема продаж товара, необходимых для построения в будущем факторной модели прогноза прибыли. В процессе исследования использовались методы обработки данных, такие как фильтрация, работа с пропущенными, аномальными и зашумленными значениями, выявление зависимых факторов. Для фильтрации данных применен алгоритм Левенштейна как один самых распространенных для сравнения строк. На реальных данных демонстрируется эффективность реализованных алгоритмов с точки зрения перспектив ее реализации в современных системах управления торговлей. Рассмотренный пример реализации обработки данных можно применить для произвольного товарного ряда других предприятий сферы розничной и оптовой торговли за счет неизменности архитектуры разработанного инструмента анализа. Это может значительно удешевить и упростить анализ данных для среднего бизнеса, а также сбор, хранение и анализ статистики.

Ключевые слова: *алгоритм Левенштейна, анализ продаж, обработка данных, статистические данные, товарный ряд, факторная модель, фильтрация данных.*

У статті розглянуто приклад ведення та аналізу статистики продажів компанії з метою формування на його основі та реалізації алгоритмів підготовки даних для ефективної роботи моделей продажів. Виділено чинники, що впливають на зміну обсягу продажів товару, необхідних для побудови в майбутньому факторної моделі прогнозу прибутку. У процесі дослідження використовувалися методи обробки даних, такі як фільтрація, робота з пропущеними, аномальними значеннями, виявлення залежних чинників. Для фільтрації даних застосовано алгоритм Левенштейна як один із найбільш розповсюджених для порівняння рядків. На реальних даних демонструється ефективність реалізованих алгоритмів із погляду перспектив її реалізації в сучасних системах управління торгівлею. Розглянутий приклад реалізації обробки даних можливо застосувати для довільного товарного ряду інших підприємств сфери роздрібної та оптової торгівлі за рахунок незмінності архітектури розробленого інструменту аналізу. Це може значно здешевити і спростити аналіз даних для середнього бізнесу, а також збір, зберігання та аналіз статистики.

Ключові слова: *алгоритм Левенштейна, аналіз продажів, обробка даних, статистичні дані, товарний ряд, факторна модель, фільтрація даних.*

The article considers an example of maintaining and analyzing statistics of company sales. This issue plays an important role in modeling, as the parameters of the model are completely determined by the data. The paper aims at conducting an analysis of the relevant statistics, forming following it and realizing the algorithm of data pre-processing for the effective performance of sales patterns. The authors analyze the shortcomings of the automated accounting system used in the company from the point of view of entering and storing statistical data that flow in the company. The factors affecting the change in sales volume of goods necessary to build a factor model of profit forecast in future are identified. The statistics, which are used in this paper, are collected for the period 2015 – 2019. The data volume is 51510 records. Since the data structure is factor one in its general form, and one cannot approximate such data, the records have missing values. It is shown that the renewal of missing values by virtue of traditional approximation methods is impossible, taking into account the specifics of the sample under consideration. The effectiveness of the implemented algorithms for filtering, parsing and binding of the input data in terms of the prospects for its implementation in the modern trade management systems is demonstrated on the basis of actual data. To solve the problems of the incoherence of input data, the requests originated from the automated accounting system are analyzed to convert the data into a single format for their subsequent binging on sales and product returns. The concern of the heterogeneity of data is handled by filtering that allows the authors to identify similar records related to the same object and combine them. The authors use the Levenshtein algorithm, as one of the most widespread for strings comparison, for filtering data. The article describes a tool which is implemented in C# programming language for working with dictionaries. The presented data processing is applicable to an arbitrary product range of other retail and wholesale enterprises due to the firmness of the architecture of the developed analysis tool. The above can significantly reduce the price and simplify data analysis for medium-sized businesses as well as the collection, storage and analysis of statistics.

Key words: *Levenshtein algorithm, sales analysis, data processing, statistics, product line, factor model, data filtering.*

Постановка проблемы в общем виде и ее связь с важными научными и практическими заданиями. В настоящее время получение достоверной информации, ее быстрый и эффективный анализ стали важнейшими предпосылками успешной финансовой деятельности любого коммерческого предприятия. Для объективной оценки эффективности продаж большое значение имеет возможность математического моделирования и анализа имеющихся статистических данных с целью прогнозирования, анализа и планирования объема продаж. Имея математическую модель по продажам, намного проще разработать маркетинговую стратегию развития компании, проводить грамотную ассортиментную политику и разрабатывать эффективные трейд-маркетинговые мероприятия [1].

Отправной точкой в процессе моделирования являются данные, характеризующие исследуемый объект, подготовить и систематизировать которые – отдельная задача. Поскольку параметры модели полностью определяются исходными данными, требуется тщательный подход к их качеству. Ошибочные, аномальные и зашумленные данные могут привести к моделям и выводам, не имеющим отношения к действительности, поэтому вопрос анализа данных играет важную роль в моделировании. Необходимой является предобработка данных.

Анализ последних исследований и публикаций, в которых положено начало решению данной проблемы и на которые опираются авторы. При обработке данных существует много ограничений, на которые нужно адекватно реагировать и выбирать правильный путь их преобразования и выравнивания. Анализ и подбор методов для качественной оценки обрабатываемых данных, работа с пропущенными значениями при факторном анализе, а также дисперсионная оценка влияния разных факторов рассматриваются в [2–5], и демонстрируется основная семантика работы с данными, аналогичными рассматриваемым в статье.

Для решения задачи интеграции разнородных источников возникает необходимость сопоставления, согласования и объединения различных представлений данных, а также исключения дублирующейся информации. На данный момент существует достаточно большое число публикаций, посвященных проблеме дублирующихся записей. В подавляющей части работ для сравнения строки используется стандартная метрика Левенштейна [6–8].

Выявление таких дубликатов широко используется при обнаружении заимствований в текстовых документах [9; 10].

Формулирование целей статьи **(постановка задания)**. Целью работы является проведение анализа актуальной статистки, формирование и реализация на его основе алгоритмов подготовки данных для эффективной работы моделей продаж.

Изложение основного материала исследования с полным обоснованием полученных научных результатов. Задача подготовки и систематизации данных для дальнейшего моделирования проводилась на реальных данных о продажах ООО «Данлер», специализирующегося на реализации обуви и сопутствующих товаров с широким ассортиментом. Компания осуществляет свою деятельность с 2003 г. Автоматизированная система бухгалтерского учета установлена только с 2015 г., и на тот момент

занесение информации не было приоритетной задачей, так как в компании не планировался сбор данных для дальнейшей обработки. Стоит отметить важный аспект анализа данных в пределах среднего бизнеса и его неподготовленности к работе с новыми технологиями. Основными проблемами являются небрежность хранения данных, а также некорректность заполнения и введение автоматизированной системы бухгалтерского учета в компании. Данные вносились разными сотрудниками без специальной подготовки, отсутствовал единый подход к заполнению и структуре одной и той же информации. В связи с указанным в первичных данных присутствовала неоднородность, что делало анализ невозможным.

Выявлена необходимость предобработки данных: фильтрация и парсинг, восстановление правильной структуры и семантики данных.

Статистика, использованная в данном исследовании, собиралась за период 2015–2019 гг. Объем данных составил 51 510 записей. Стоит отметить, что в этих записях остаются пропущенные значения, которые некорректны для анализа, так как структура данных в своем общем виде является факторной и функционально аппроксимировать такие значения нельзя. Анализ проводится на ежемесячной и ежеквартальной основе.

В табл. 1 приведен перечень выделенных необходимых показателей анализа продаж.

Одной из основных проблем при извлечении данных было то, что бухгалтерская система, установленная на предприятии, не позволяла выбирать данные по одной единице (каждой товарной позиции), а только сгруппированные товары в специальном отчете. Проблему усугубил факт того, что товары по приходу, продаже и возврату не были связаны. Поэтому запросы, поступавшие из автоматизированной бухгалтерской системы, были изъяты в базу с помощью профайлера и преобразованы в единственный формат, а затем по ключу связаны с продажами и возвратами. При связывании данных была обнаружена небольшая погрешность (не более 1%), которой относительно общего объема данных можно пренебречь. Несвязанные данные были удалены.

Изначально данные, а именно описательная часть товара, были представлены в виде одной строки, что затруднило их дальнейший анализ. Средствами языка C# реализованы фильтры для работы со словарями, которые выбирали корректные значения для каждого поля, представленного в табл. 1. Входящими значениями являлся картеж из товара, который проходил парсинг и сохранялся в результирующую таблицу. По атрибуту «количество товара» кортеж копировался n раз. Далее проходила фильтрация данных в кортеже и вставка уже нормализованных факторов. Фильтрация данных по словарю осуществлялась по алгоритму Левенштейна.

Применение фильтров позволило: привести записи, полученные из разнородных источников, к единой схеме данных; выявить похожие записи, относящихся к одному и тому же объекту, и объединить их с содержанием всех соответствующих атрибутов без избыточности.

Выводы из этого исследования и перспективы дальнейших исследований в этом направлении. Представленные в статье результаты анализа реальных данных демонстрируют реализуемость и эффективность использования предложенных алгоритмов с точки зрения перспектив ее реализации в современных системах управления торговлей с целью расширения их функциональности. Анализ данных расположен на стороне базы данных в виде процедур, то есть аппарат анализа может расширяться, что означает постоянное улучшение отчетов и ответов на поставленные вопросы с помощью статистики.

Разработанный инструмент анализа данных является индивидуальным относительно ком-

Таблица 1

Описание структуры данных

Показатель	Комментарии
Фирма-производитель	Сбор статистики по продажам фирмы-производителя может дать ответ на ключевой вопрос релевантности выбора товара покупателем и в итоге оптимизировать выбор партнеров для следующего закупочного сезона. Факторные данные: 428 уникальных поставщиков. Тип данных: строка.
По прайсу	Фактическая цена товара на момент поступления. Непрерывная величина. Тип данных: double.
Себестоимость единицы продукции	Себестоимость товара является важным аспектом любого анализа продаж. Зная уровень себестоимости продукта, проще разрабатывать трейд-маркетинговые акции и управлять ценообразованием в компании. На основе себестоимости можно рассчитать среднюю рентабельность продукта и определить наиболее выгодные с точки зрения прибыли позиции для стимулирования продаж. Непрерывная величина. Тип данных: double.
Дата и время поступления на склад	Дата и время поступления на склад. Тип данных: datatime.

Закінчення табл. 1

Категория товара	Категория представленной в товарном ряде обуви; факторные данные: 1 «Сапоги» 2 «Ботинки» 3 «Туфли» 4 «Бальзам» 5 «Балетки» 6 «Средства для ухода за обувью» 7 «Кросовки» 8 «Шлепанцы» 9 «Босоножки» 10,NA (Null Attribute) 11 «Мокасины» 12 «Спортивная обувь» 13 «Тапки» 14 «Сабо» 15 «Сандалии» 16 «Сумка» 17 «Слипоны» 18 «Угги» 19 «Ремень» 20 «Полуботинки» 21 «Снегоступы» 22 «Кеды» 23 «Ботильйоны» 24 «Дисконтная карта» 25 «Носки» 26 «Подарочный сертификат» Тип данных: строка.
Детское	Факторный тип данных, который принимает два значения: да или нет. Тип данных: строка.
Сезон	Факторный тип данных, который принимает три значения: лето, зима, демисезонное. Тип данных: строка.
Пол	Факторный тип данных, который принимает два значения: м или ж. Тип данных: строка.
Размер	Факторный тип данных, который принимает значения натуральных чисел в диапазоне 34–48. Тип данных: integer.
Страна-производитель	Факторный тип данных, который принимает значения: 1 «Украина» 2 «Польша» 3 «Италия» 4 NA (Null Attribute) 5 «Турция» 6 «Германия» 7 «Китай» 8 «Молдова» 9 «Испания» 10 «Великобритания» 11 «Португалия» 12 «Голландия» 13 «США» 14 «Бразилия» Тип данных: строка.
Покупатель	Факторный тип данных, ФИО покупателя, если покупатель не зарегистрирован – «Покупатель». Тип данных: строка.
Фактическая цена	Цена товара на момент продажи товара. Непрерывная величина. Тип данных: double.
Продавец	Факторный тип данных, ФИО продавца. Тип данных: строка.
Дата и время продажи	Дата и время продажи, может быть NA (Null Attribute). Тип данных: datatime.
Дата и время возврата	Дата и время возврата товара, может быть NA (Null Attribute). Тип данных: datatime.

пании, но подход к разработке и архитектуре остается неизменным и может быть использован для анализа товарного ряда других предприятий.

Система анализа данных может быть расширена за счет выгрузки статистики из Интернет-магазина. Указанные данные имеют другую структуру, и необходимо их приведение к единой форме. Это является трудоемким и ресурсозатратным этапом, которому следует уделить больше времени при разработке системы, так как последующие результаты моделей и отчетов для анализа будут предоставляться в зависимости от исходной формы.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК:

1. Науменко Н.Ю., Короткая Л.И. Разработка информационно-аналитической системы анализа финансовой деятельности предприятий. *Економічний вісник ДВНЗ «Український державний хіміко-технологічний університет»*. 2016. № 1(3). С. 16–20.

2. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R / пер. с англ. Москва : ДМК Пресс, 2014. 588 с.

3. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. Санкт-Петербург : Питер, 2017. 336 с.

4. Бослаф С. Статистика для всех. Москва : ДМК Пресс, 2017. 586 с.

5. Наглядная статистика. Используем R! / А.Б. Шипунов и др. Москва : ДМК Пресс, 2014. 298 с.

6. Graham A. Stefen. String Searching Algorithms. Singapore : World Scientific Publishing Co. Pte. Ltd, 2000. 235 p.

7. Lutabingwa C.J. Auriacombe. Data Analysis In Quantitave Research. *Juornal of public Administration.* 2007. Vol. 42. № 6. P. 528–548.

8. Si Lhoussain A., Hicham G., Abdella Y. Adaptating The Levenshtein Distance To Contextual Spelling Correction. *International Journal of Computer Science and Applications, Technomathematics Research Foundation.* 2015. Vol. 12. № 1. P. 127–133.

9. Rakian S., Safi Esfahani F., Rastegarih. A Persian Fuzzy Plagiarism Detection Scheme Based On Semantic Rolelabeling. *Journal Of Information Systems and Telecommunication*. 2015. Vol. 3. P. 182–190.

10. Ezzikouri H., Erritali M., Oukessou M. Fuzzy-Semantic Similarity For Automatic Multilingual Plagiarism Detection. *International Journal of Advanced Computer Science and Applications*. 2017. Vol. 8. № 9. P. 86–90.

REFERENCES:

1. Nnaumenko N. yu., Korotkaya L. I. (2016) Razrabotka informatsionno analiticheskoy sistemy analiza finansovoy deyatel'nosti predpriyatiy. [Development of an information-analytical system for analyzing the financial activities of enterprises]. Ekonomichniy visnik DVNZ «Ukraïns'kiy derzhavniy khimiko-tekhnologichniy universitet», no. (3). pp.16–20.

2. Robert I. Kabakov. (2014) R v deystvii. Analiz i vizualizatsiya dannykh v programme R. [R in action. Analysis and visualization of data in the program R] Moscow. (in Russian)

3. Silen D., Meysman A., Ali M. (2017) Osnovy Data Science i Big Data. Python i nauka o dannykh. [Fundamentals of Data Science and Big Data. Python and data science]. Piter : SPb. (in Russian)

4. Boslaf S. (2017) Statistika dlya vsekh. [Statistics for everyone]. Moscow : DMK Press. (in Russian)

5. Shipunov A.B. (2014) Naglyadnaya statistika. Ispol'zuem R! [Visual statistics. Use R!] / Shipunov A.B. i dr. Moscow : DMK Press. (in Russian)

6. Graham A. Stefen. String Searching Algorithms. Singapore : World Scientific Publishing Co. Pte. Ltd, 2000. 235 p.

7. J. Lutabingwa, C. J. Auriacombe. Data Analysis In Quantitave Research. *Juornal of public Administration.* 2007. Vol. 42. № 6. P. 528–548.

8. A. Si Lhoussain, G. Hicham, Y. Abdella. Adaptating The Levenshtein Distance To Contextual Spelling Correction. *International Journal of Computer Science and Applications, Technomathematics Research Foundation.* 2015. Vol. 12. № 1. P. 127–133.

9. Rakian S., Safi Esfahani F., Rastegarih. A Persian Fuzzy Plagiarism Detection Scheme Based On Semantic Rolelabeling. *Journal Of Information Systems and Telecommunication*. 2015. Vol. 3. P. 182–190.

10. Ezzikouri H., Erritali M., Oukessou M. Fuzzy-Semantic Similarity For Automatic Multilingual Plagiarism Detection. *International Journal of Advanced Computer Science and Applications*. 2017. Vol. 8. № 9. P. 86–90.