Для цитирования: Вершинин Е. В., Прокофьев М. Л., Афанасьев В. Р. Проектирование аналитической системы обработки фискальных данных // Вопросы радиоэлектроники. 2019. № 3. С. 78–82.
DOI 10.21778/2218-5453-2019-3-78-82
УДК 004.67

Е. В. Вершинин[1], М. Л. Прокофьев[1], В. Р. Афанасьев[1]

[1] Московский государственный технический университет им. Н. Э. Баумана, Калужский филиал

ПРОЕКТИРОВАНИЕ АНАЛИТИЧЕСКОЙ СИСТЕМЫ ОБРАБОТКИ ФИСКАЛЬНЫХ ДАННЫХ

В работе рассматривается задача проектирования аналитической системы, предназначенной для обработки фискальных данных. С точки зрения бизнеса, такая система должна решать задачу анализа рыночной корзины, то есть поиска наиболее типичных шаблонов покупок. С точки зрения интеллектуального анализа данных, решается задача поиска ассоциативных правил, состоящая из двух этапов: поиска всех частых наборов с их значениями поддержки и получения ассоциативных правил на основе найденных наборов. Первый этап обеспечивается различными алгоритмами поиска частых наборов. В работе в качестве оптимального выбран алгоритм Frequent Pattern Growth Strategy (FPG). Приводится математическая формулировка поставленной задачи и метод реализации выбранного алгоритма в рамках целевой системы. Результатом работы является описание отказоустойчивой и масштабируемой модели аналитической системы.

Ключевые слова: *интеллектуальный анализ данных, алгоритм поиска частых наборов, Frequent Pattern Growth*

Введение

В настоящее время вопросы интеллектуального анализа данных (Data Mining) становятся все более актуальными. Это связано с тем, что с развитием информационных технологий процессы сбора и хранения данных стали существенно проще и дешевле. Для того чтобы извлекать из накопленных данных полезную информацию, необходимо применять соответствующие алгоритмы обработки информации. Разработка таких алгоритмов требует значительных интеллектуальных усилий.

Операторы фискальных данных (ОФД) предоставляют своим клиентам инструменты анализа фискальных данных. Однако перечень таких инструментов в большинстве случаев ограничивается просмотром объемов выручки, прошедших через ОФД, а также общей статистики по конкретным кассам (выручка, количество чеков и т.п.).

Таким образом, можно сделать вывод, что данные, которыми обладают ОФД, используются недостаточно эффективно. Повысить эффективность можно за счет применения современных методов интеллектуального анализа. Например, можно строить математические модели, позволяющие прогнозировать какие-либо события (изменения спроса на определенный товар, отток покупателей и т.п.), или же выявлять в данных различные нетривиальные паттерны (в каком районе выше спрос на определенные товары или наличие каких товаров влияет на спрос на другие товары). Такая информация может помочь клиентам ОФД повысить эффективность своих продаж и, как следствие, получить дополнительную прибыль.

Цель работы – создание модели аналитической системы обработки фискальных данных, способной успешно решать обозначенные задачи.

Решение задачи анализа рыночной корзины на основе имеющихся фискальных данных

Под анализом рыночной корзины будем понимать процесс поиска наиболее типичных шаблонов покупок в супермаркетах путем анализа баз данных транзакций с целью определения комбинаций товаров, связанных между собой [1]. Полученные в ходе анализа рыночной корзины результаты могут быть использованы для решения следующих задач:

- оптимизация ассортимента и его размещения в торговых залах;
- повышение эффективности управления запасами;
- увеличение объемов продаж за счет предложения клиентам сопутствующих товаров;
- оценка эффективности различных рекламных кампаний (промоакций);
- формирование персональных рекомендаций.

С точки зрения интеллектуального анализа данных, анализ рыночной корзины является задачей поиска ассоциативных правил [2]. Рассмотрим ее математическую формулировку.

Пусть $\chi = \{x1, x2, ..., xm\}$ – множество значений, называемых элементами. Множество $X \subseteq \chi$ называется набором. Набор мощности (или размера) k называется k-набором. Через $\chi^{(k)}$ обозначим множество всех k-наборов, то есть подмножеств χ с размером k.

Пусть $\tau = \{t1, t2, ..., tn\}$ – множество других элементов, называемых идентификаторами транзакций, или TIDs. Множество $T \subseteq \tau$ называется набором идентификаторов транзакций (TID-набором). Предположим, что наборы и TID-наборы хранятся в упорядоченном виде.

Транзакцией называется кортеж формы $\langle t, X \rangle$, где $t \in \tau$ – уникальный идентификатор транзакции, а X – набор элементов. Набор транзакций τ может обозначать набор всех покупок клиентов в супермаркете.

Бинарная база данных D представляет собой отношение между TID-набором и набором элементов в виде $D \subseteq \tau \times \chi$.

На рис. 1 приведен пример двоичной базы данных для базы данных транзакций. Здесь $\chi = \{A, B, C, D, E\}$ и $\tau = \{1,2,3,4,5,6\}$. В двоичной базе данных ячейка в строке t и столбце x равна 1, только когда $(t, x) \in D$, а в противном случае равна 0.

Поддержка набора элементов X в базе данных D представляет собой количество транзакций в D, которые содержат X, и рассчитывается по формуле

$$\sup(X, D) = \left\{ t \middle| \langle t, i(t) \rangle \in D \text{ и } X \subseteq i(t) \right\} = |t(x)|. \quad (1)$$

Считается, что набор элементов X является частым в D, если $\sup(X, D) \geqslant \text{minsup}$, где minsup – минимальный порог поддержки, устанавливаемый пользователем.

Ассоциативное правило представляет собой выражение $X \overset{s, c}{\rightarrow} Y$, где X и Y являются непересекающимися наборами, то есть $X, Y \subseteq \chi$ и $X \cap Y = \varnothing$.

Поддержка правила – количество транзакций, в которых одновременно находятся оба набора X и Y, рассчитывается по формуле

$$\sup(X \rightarrow Y) = |t(XY)| = \sup(XY). \quad (2)$$

Достоверность правила – условная вероятность того, что транзакция содержит Y, учитывая, что она содержит X, рассчитывается по формуле

$$\text{conf}(X \rightarrow Y) = P(Y|X) = \frac{P(X \wedge Y)}{P(X)} = \frac{\sup(XY)}{\sup(X)}. \quad (3)$$

Ассоциативное правило является частым, если набор XY является частым, то есть $\sup(XY) \geqslant \text{minsup}$, и является надежным, если $\text{conf} \geqslant \text{minconf}$, где minconf – заданный пользователем минимальный порог доверия [3].

Исходя из определений поддержки и достоверности ассоциативных правил, можно сказать, что задача поиска ассоциативных правил состоит из двух этапов: поиска всех частых наборов с их значениями поддержки и поиска всех частых ассоциативных правил, удовлетворяющих критерию надежности.

На первом этапе необходимо рассмотреть существующие алгоритмы поиска частых наборов [4]. Классическим алгоритмом для решения такого рода задач является Apriori, однако, как и большинству алгоритмов поиска частых наборов, для получения результата ему необходимо полностью обходить базу данных транзакций. Учитывая тот факт, что данные о новых транзакциях будут регулярно поступать от ОФД, применение такого подхода нерационально, поскольку полное сканирование базы данных может занимать большое количество времени [5].

Альтернативой для такого подхода является алгоритм Frequent Pattern Growth Strategy (FPG). В его основе лежит предобработка базы данных транзакций, при которой она преобразуется в компактную древовидную структуру, называемую Frequent Pattern Tree – префиксное дерево частых наборов (FP-дерево). Соответственно, и алгоритм логически разделен на два этапа: построение FP-дерева; извлечение частых наборов из FP-дерева.

Данные этапы могут выполняться независимо друг от друга в различные моменты времени. Таким образом можно постепенно собирать данное

t	X
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

а)

TID	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1;	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

б)

Рисунок 1. Представление базы данных транзакций: а – БД транзакций; б – бинарная БД

префиксное дерево и выполнять второй этап алгоритма именно в тот момент, когда в этом появляется потребность. Кроме того, благодаря своей компактной структуре префиксное дерево можно кешировать, тем самым повышая скорость обработки запросов к системе.

Для того чтобы построенное дерево имело как можно более компактную и сбалансированную структуру, необходимо добавлять в него элементы транзакций в упорядоченном виде (в порядке убывания поддержки). Однако вышеописанный подход не позволяет полностью выполнять это условие, поскольку с течением времени поддержка отдельных элементов и соответственно схема упорядочивания элементов в транзакциях могут меняться. Частично компенсировать данный недостаток можно, осуществляя накопление транзакций и обновляя дерево не при регистрации каждой новой транзакции, а при возникновении специального события, например по истечении заданного временного интервала. Это позволит избежать несбалансированного построения дерева в ситуациях, когда на малом промежутке времени рост поддержки отдельных элементов является несистемным. Для того чтобы кардинально решить проблему несбалансированности дерева, можно полностью перестраивать его через определенные интервалы времени, однако, для этого необходимо реализовать сохранение данных о получаемых транзакциях в БД.

В соответствии с вышесказанным спроектирована модель аналитической системы, приведенная на рис. 2.

На данной схеме приведены функциональные модули, а также их связи друг с другом и с внешними по отношению к системе сущностями. Источник данных (транзакций), обозначенный в схеме как ОФД, связан с модулем предварительной обработки не напрямую, а посредством очереди сообщений. Данный подход позволит, во-первых, оградить модуль предварительной обработки от перегрузки, во-вторых, избежать потери данных, если модуль обработки по какой-либо причине будет недоступен, например, во время обновления системы.

При получении новой транзакции модуль предварительной обработки будет производить ее обработку следующим образом: извлекать из транзакции требуемые данные, т.е. список элементов (товаров), участвующих в транзакции, преобразовывать эти данные к внутреннему формату системы, после чего записывать эту информацию в БД с сохранением времени записи. На основе сохраненных транзакций будет строиться специальная структура данных, которая позволит эффективно извлекать аналитическую информацию, впоследствии с ней будет работать модуль анализа. Однако с течением времени данная структура станет неактуальной, так как в систему продолжат поступать новые транзакции, и, соответственно, созданную ранее структуру данных необходимо будет актуализировать. Для этого во время работы системы будет запущена специальная служба, которая через определенные промежутки времени будет извлекать из БД транзакции, фильтруя их по временному признаку, и дополнять текущую структуру новыми данными. Таким образом, модуль анализа будет получать последнюю актуальную версию данных из БД и кешировать ее для исключения задержки на обращение к БД при выполнении пользовательского запроса.

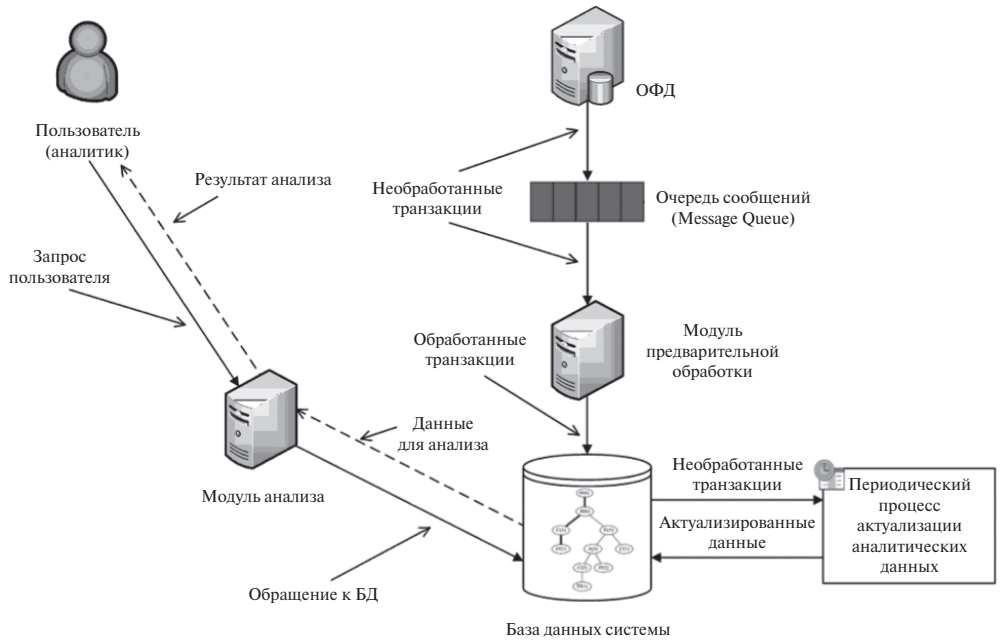

Рисунок 2. Модель системы

Основную нагрузку в данной системе будет испытывать модуль анализа, поэтому для повышения горизонтальной масштабируемости имеет смысл распараллеливать сложные вычислительные процессы [6].

Выводы

С целью проектирования аналитической системы, предназначенной для обработки фискальных данных, рассмотрена математическая формулировка задачи поиска ассоциативных правил. Установлено, что решение указанной задачи состоит из двух этапов: поиска всех частых наборов с их значениями поддержки и поиска всех частых ассоциативных правил, удовлетворяющих критерию надежности. Для реализации первого этапа предложено использовать алгоритм Frequent Pattern Growth Strategy, основанный на предварительной обработке базы данных транзакций, при которой она преобразуется в компактную древовидную структуру. Предложена модель отказоустойчивой и масштабируемой аналитической системы, предназначенной для обработки фискальных данных с целью анализа рыночной корзины для поиска наиболее типичных шаблонов покупок.

СПИСОК ЛИТЕРАТУРЫ

1. Анализ рыночной корзины [Элекронный ресурс]. URL: https://basegroup.ru/community/glossary/market-basket (дата обращения: 10.12.2018).
2. Maimon O., Rokach L. Data mining and knowledge discovery handbook. 2nd edition. Springer Science+Business Media, 2010. 1285 p.
3. Zaki M.J., Meira W. Data mining and analysis. Cambridge University Press, 2014. 593 p.
4. Charu C. Aggarwal, Han J. Frequent pattern mining. Springer, 2014. 471 p.
5. Han J., Pei J., Yin Y., et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach // Data Mining and Knowledge Discovery. 2004. Vol. 8. Iss. 1. P. 53–87.
6. Li H., Wang Y., Zhang D., et al. PFP: parallel FP-growth for query recommendation. Proceedings of the 2008 ACM conference on Recommender systems, 2008. P. 107–114.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Вершинин Евгений Владимирович, к.ф.-м.н., доцент, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (Национальный исследовательский университет)», Российская Федерация, 248000, Калуга, ул. Баженова, д.2, тел.: 8 (4842) 74-05-95, e-mail: yevgeniyv@mail.ru.

Прокофьев Михаил Львович, магистрант, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (Национальный исследовательский университет)», Российская Федерация, 248000, Калуга, ул. Баженова, д.2, тел.: 8 (910) 510-21-63, e-mail: mp.prokofyev@gmail.com.

Афанасьев Владислав Романович, магистрант, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (Национальный исследовательский университет)», Российская Федерация, 248000, Калуга, ул. Баженова, д.2, тел.: 8 (962) 172-78-10, e-mail: ya.vladicl@ya.ru.

For citation: Vershinin Ye.V., Prokofyev M.L., Afanasyev V.R. Developing fiscal data processing analytical system. Voprosy radioelektroniki, 2019, no. 3, pp. 78–82.
DOI 10.21778/2218-5453-2019-3-78-82

Ye.V. Vershinin, M.L. Prokofyev, V.R. Afanasyev

DEVELOPING FISCAL DATA PROCESSING ANALYTICAL SYSTEM

The paper deals with the task of designing an analytical system for processing fiscal data. From a business point of view, such a system should solve the problem of analyzing a market basket, that is, finding the most typical patterns of purchases. From the point of view of data mining, the task of searching for association rules is solved, which consists of two stages: the search for all frequent sets with their support values and the acquisition of association rules based on the sets found. The first stage is provided by various search algorithms for frequent sets. In the paper, the algorithm chosen is the Frequent Pattern Growth Strategy (FPG) as the optimal one. The mathematical formulation of the task and the method for implementing the selected algorithm within the target system are given. The result of the work is a description of the fault-tolerant and scalable model of the analytical system.

Keywords: frequent set search algorithms, intellectual data processing, Frequent Pattern Growth

REFERENCES

1. Market basket analysis. Available at: https://basegroup.ru/community/glossary/market-basket (accessed 10.12.2018). (In Russian).
2. Maimon O., Rokach L. *Data mining and knowledge discovery handbook.* 2nd edition. Springer Science+Business Media, 2010, 1285 p.

3. Zaki M.J., Meira W. *Data mining and analysis.* Cambridge University Press, 2014, 593 p.
4. Charu C. Aggarwal, Han J. *Frequent pattern mining.* Springer, 2014, 471 p.
5. Han J., Pei J., Yin Y., et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. *Data Mining and Knowledge Discovery,* 2004, vol. 8, iss. 1, pp. 53–87.
6. Li H., Wang Y., Zhang D., et al. PFP: parallel FP-growth for query recommendation. Proceedings of the 2008 ACM conference on Recommender systems, 2008, pp. 107–114.

AUTHORS

Vershinin Yevgeniy, Ph. D., associate professor, Bauman Moscow State Technical University (Kaluga Branch), 2, Bazhenova St., Kaluga, 248000, Russian Federation, tel.: +7 (4842) 74-05-95, e-mail: yevgeniyv@mail.ru.
Prokofyev Mikhail, master, Bauman Moscow State Technical University (Kaluga Branch), 2, Bazhenova St., Kaluga, 248000, Russian Federation, tel.: +7 (910) 510-21-63, e-mail: mp.prokofyev@gmail.com.
Afanasyev Vladislav, master, Bauman Moscow State Technical University (Kaluga Branch), 2, Bazhenova St., Kaluga, 248000, Russian Federation, tel.: +7 (962) 172-78-10, e-mail: ya.vladicl@ya.ru.