ПРИМЕНЕНИЕ СИСТЕМЫ DEEP DATA DIVER ДЛЯ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА РЫНОЧНЫХ КОРЗИН

М. Г. Асеев[1], В. А. Дюк[2]

[1]Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия В.О., д.39
<maxim@datadiver.nw.ru>
[2]Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия В.О., д.39
<duke@datadiver.nw.ru>

УДК 681.3
М. Г. Асеев, В. А. Дюк. **Применение системы Deep Data Diver для решения задачи анализа рыночных корзин** // Труды СПИИРАН. Вып. 2, т. 1. — СПб.: СПИИРАН, 2004.
Аннотация. *Система Deep Data Diver использует новую технологию поиска ассоциативных правил, основанную на модифицированном аппарате линейной алгебры с использованием процедуры самоорганизации данных и эффекта информационного структурного резонанса. Уникальные свойства системы позволяют находить в данных высокоточные ассоциации элементов исходного множества транзакций с заданным элементом. Эти множества образуют корзину с высоким уровнем обеспечения (support) и длинным набором элементов (long itemsets). В статье дается общая характеристика системы Deep Data Diver и приводятся сравнительные результаты решения конкретной задачи анализа рыночных корзин.* — Библ. 6 назв.

UDC 681.3
M. G. Aseev, V. A. Duke. **Deep Data Diver Application in Market Baskets Analysis** // SPIIRAS Proceedings. Issue 2, vol. 1. — SPb.: SPIIRAS, 2004.
Abstract. *The Deep Data Diver system uses a new technology of associative rules search which is based on modified tools of linear algebra and the usage of a data self-organization procedure and an informational structural resonance effect. The unique characteristics of the system allow to search data for highly accurate associations of the items comprising the initial transaction set with a given item. These sets form a basket with high support level and long itemsets. The article provides a general overview of the Deep Data Diver system and gives the comparison results of solving the specific task of market basket analysis.* — Bibl. 6 items.

1. Введение

Рассмотрим классическую формулировку задачи анализа рыночных корзин [1]. Рыночная корзина — это набор товаров, приобретенных покупателем в рамках одной отдельной транзакции. Сюда относятся, например, результаты визита покупателя в бакалейную лавку, интерактивная покупка в виртуальном магазине типа Amazon.com и пр. Регистрируя бизнес-операции в течение всего времени своей деятельности, торговые компании накапливают огромные собрания таких транзакций (базы данных).

Одна из наиболее распространенных задач статистического анализа подобных баз данных, состоит в поиске товаров или наборов товаров (itemset), которые одновременно встречаются во многих транзакциях. Шаблоны поведения покупателей, выявленные благодаря такому анализу, в общем виде характеризуются перечнем входящих в набор товаров и числом транзакций, содержащих эти наборы. Торговые компании используют эти шаблоны для того, чтобы более правильно разместить товары в магазинах, изменить структуру страниц товарных каталогов и Web страниц и т.п.

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Набор, состоящий из i товаров, называется i-элементным набором (i-itemset). Процент транзакций, содержащих данный набор, называется «обеспечением» (support) набора. Принято считать, что для того чтобы набор представлял интерес, его обеспечение должно быть выше определенного пользователем минимума; такие наборы называют часто встречающимися (frequent).

Для набора товаров часто используют характеристику «доверие» (confidence), связанную с точностью выявления набора тем или иным алгоритмом. Точность всегда определяется по отношению к одному из элементов набора. Она равна вероятности того, что при обязательном вхождении в набор $(i-1)$ элементов в него войдет также некоторый i-й элемент. Чем выше «доверие» у выделенного набора, тем более значимым для реальной практики является рассматриваемый набор.

Кроме того, важной характеристикой является длина набора i. Проблеме поиска в данных длинных наборов (long itemsets) является весьма актуальной. Она достаточно подробно рассматривается в работе [2].

Одним из первых алгоритмов анализа рыночной корзины был алгоритм Apriori [3]. Данный алгоритм определяет часто встречающиеся наборы за несколько этапов. Каждый этап состоит из двух шагов: формирование кандидатов (candidate generation) и подсчет кандидатов (candidate counting). На 1-м этапе выбранное множество наборов-кандидатов содержит все 1-элементные наборы. Алгоритм вычисляет их обеспечение во время шага подсчета кандидатов. Затем Apriori сокращает множество наборов-кандидатов отсеивая тех кандидатов, которые не могут быть часто встречающимися. Отсев происходит на основе простого предположения о том, что если набор является часто встречающимся, все его подмножества также должны быть часто встречающимися. 2-й и последующие этапы используют аналогичные операции по отношению к наборам из 2-х, 3-х и т.д. элементов.

В настоящее время известно большое количество Apriori-подобных алгоритмов. Основным недостатком этих алгоритмов считается их неспособность находить длинные наборы. Поэтому исследователи предпринимали и предпринимают попытки изобрести новые подходы и алгоритмы для анализа рыночных корзин. Достаточно подробно указанные попытки описаны в [2], где акцент делается на свойстве масштабируемости алгоритмов (линейной зависимости времени поиска решения от размера анализируемой базы данных).

На наш взгляд, с учетом упомянутого выше аналитического обзора [2], проблема поиска длинных наборов данных за приемлемое время до сих пор не нашла своего решения. Требуются принципиально новые подходы к поиску ассоциативных правил в базах данных.

2. Основные характеристики системы Deep Data Diver™

Нами разработана система **Deep Data Diver**, которая претендует дать ответ на некоторые нерешенные вопросы. Система использует новые принцип и технологию поиска логических закономерностей в данных. Принцип основывается на представлениях специальной *локальной геометрии*. В этой геометрии каждый многомерный объект существует в собственном локальном пространстве событий с индивидуальной метрикой. За счет свойств локальных пространств комбинаторная проблема поиска логических закономерностей

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

получает геометрическое истолкование. Технология такого поиска основывается на модифицированном аппарате линейной алгебры с использованием процедуры самоорганизации данных и эффекта информационного структурного резонанса. Основные положения технологии изложены в [4, 5].

Система **Deep Data Diver** имеет следующие основные характеристики:

1. Нахождение «лучших» (наиболее полных при заданной точности) if-then правил для каждой записи базы данных.
2. Построение и тестирование классификаторов данных на основе if-then правил.
3. Построение «нечетких» if-then правил.
4. Построение дендрограмм и исследование метаструктуры множества правил.

Кроме того, важными являются дополнительные характеристики системы:

5. Линейная зависимость времени работы алгоритма поиска от объема данных.
6. Отсутствие ограничений на тип данных.
7. Работа в условиях любого количества пропусков в данных.
8. Работа в условиях «засоренных» данных.
9. Использование приема «данные + шум», способствующего выявлению устойчивых закономерностей в данных.
10. Нахождение непериодических шаблонов сложной формы в числовых и символьных рядах.
11. Возможность распараллеливания процесса поиска if-then правил.

3. Характеристика иллюстративных данных

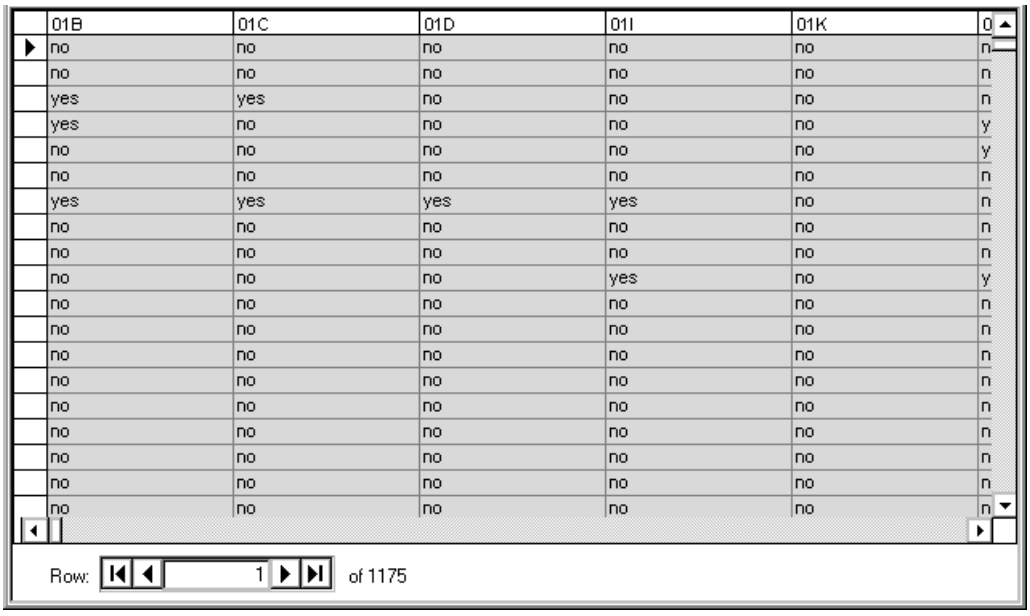
Для иллюстрации возможностей системы **Deep Data Diver** мы воспользуемся примером, опубликованном в руководстве по работе с системой PolyAnalyst [6]. В этом примере рассматривается конкретная компания – производитель электронных приборов и их компонент, выпускающая более 250 продуктов, каждый из которых маркируется трехзначным кодом. Цель нашего анализа состоит в выяснении вопроса, какие из выпускаемых продуктов продаются вместе. Часть анализируемой таблицы данных приведена на рис. 1.

В таблице на рис. 1 каждая строка представляет транзакцию, а каждая колонка соответствует коду какого-либо продукта. Всего колонок — 255. Ячейки таблицы содержат значения “yes” или “no” в зависимости от того, совершена или нет покупка того или иного товара в определенной транзакции. Таблица содержит описание 1175 транзакций. Обычно в одной транзакции совершается порядка дюжины покупок.

3.1. Результаты анализа PolyAnalyst

Прежде чем приступить к результатам анализа данных системой **Deep Data Diver**, рассмотрим решение, предоставляемое системой PolyAnalyst. Разработчики этой системы обращают внимание на превосходные свойства своего продукта. Это касается как быстродействия системы (используется подход, связанный с построением деревьев решений), так и качества получаемых решений (впрочем, аналогичные утверждения можно встретить у создателей практически любого программного продукта в области Data Mining).

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru



01B	01C	01D	01I	01K	0
no	no	no	no	no	n
no	no	no	no	no	n
yes	yes	no	no	no	n
yes	no	no	no	no	y
no	no	no	no	no	y
no	no	no	no	no	n
yes	yes	yes	yes	no	n
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	yes	no	y
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	no	no	n
no	no	no	no	no	n

Рисунок 1. Фрагмент таблицы транзакций по продаже электронных приборов и их компонент

В рассматриваемом примере в системе PolyAnalyst используются следующие настроечные параметры: минимальный уровень «обеспечения» (Minimum support level) — 10 %, минимальное «доверие» (confidence) — 65 %. Ассоциации, обнаруженные системой приведены на рис. 2.

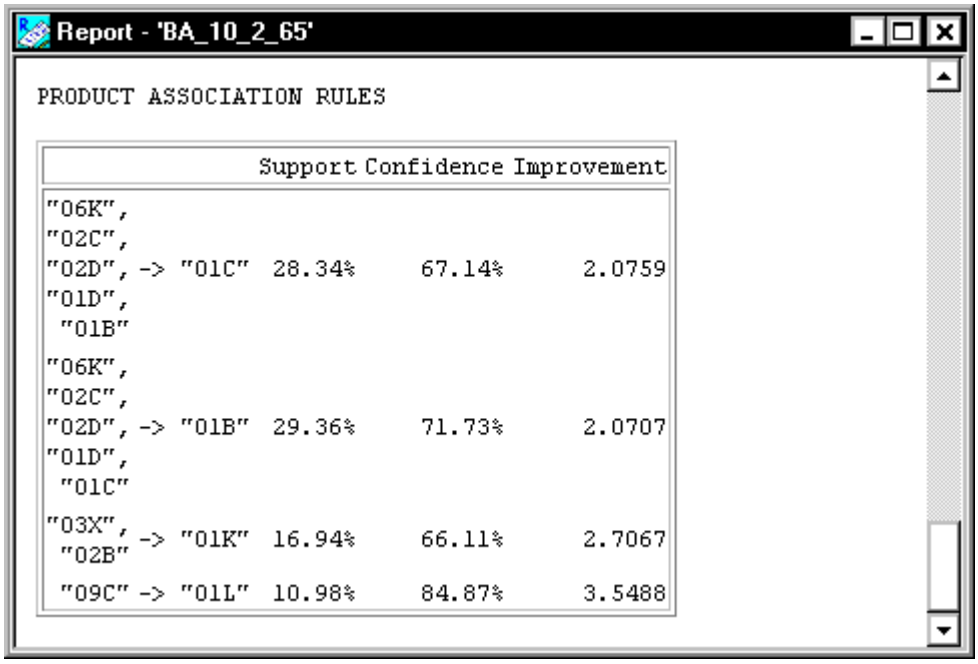


Рисунок 2. Ассоциативные правила, обнаруженные системой PolyAnalyst

Здесь следует сделать одно замечание, связанное с особенностью функционирования PolyAnalyst. Уровень обеспечения, приведенный на рис. 2, рассчитывается не по всем транзакциям, а только по тем из них, в которых встречается элемент (продукт), указанный в правой части выражения $(a_i, a_j, \ldots) \Rightarrow a_k$.

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

3.2. Результаты анализа Deep Data Diver

Начало работы в системе Deep Data Diver принципиально отличается от других систем. Здесь требуется задать 4 основных параметра:

1. Элемент (item), с которым ищутся ассоциации.
2. Номер транзакции (строку в таблице данных), для которой ищется наиболее полная при заданной точности ассоциация.
3. Желаемый минимальный уровень точности ассоциации.
4. Минимальный уровень обеспечения транзакций с заданным элементом.

В рассматриваемом примере на первом шаге работы системы мы выбираем наиболее часто покупаемый товар — 01B (407 транзакций) и наиболее насыщенную покупками транзакцию. Минимальный уровень точности ассоциации задаем равным 0,95. Получаем первую ассоциацию с продуктом 01B — **01C 01D 01I 02C ⇒ 01B.** Эта ассоциация покрывает 155 из 407 покупок товара 01B. Точность ассоциации — 0,98. Обеспечение (support — 0,38).

На следующем шаге мы снова ищем ассоциацию с тем же продуктом 01B, но изменяем номер транзакции, для которой производится поиск. Мы выбираем наиболее насыщенную покупками товара 01B транзакцию, ранее не «покрытую» первой найденной ассоциацией. Получаем вторую ассоциацию с продуктом 01B — **01C 01D 01K 02C ⇒ 01B**. Она покрывает 144 покупки товара 01B с точностью 0,99 и обеспечением 0,35. Две найденные транзакции вместе покрывают 175 покупок 01B, что составляет уже 43 %. Далее процедура продолжается аналогичным образом, для ранее не покрытых транзакций.

Всего система выявила в данных 22 ассоциации с товаром 01B, которые приведены к табл. 1.

В целом 22 ассоциации покрывают 73 % транзакций с товаром 01B, или 25 % всех транзакций, представленных в исходной таблице данных. Корзину, описываемую выявленными ассоциациями, составляют 15 товаров вместе с 01B (7 % всех товаров). В нее вошли следующие товары — 01C, 01D, 01I, 01K, 02B, 01X, 02C, 01L, 07Q, 06K, 02L, 07A, 03X, 02P и 01B. Графически найденную корзину (корзина № 1) можно представить в виде диаграммы (рис. 3). Любой из покупаемых товаров, попавших в изображенную корзину, в 73 случаях из 100 будет принадлежать к одной из 22 высокоточных ассоциаций с товаром 01B.

Однако на этом анализ данных с помощью системы Deep Data Diver не заканчивается. Далее нас начинают интересовать товары, не вошедшие в корзину № 1. Среди них, конечно, наибольший интерес представляет наиболее часто покупаемый товар. В нашем случае это товар 02D, который встречается в 228 из 1175 транзакциях.

Ассоциации с товаром 02D, обнаруженные системой, представлены в табл. 2. В корзину, описываемую этими ассоциациями вошли следующие 15 товаров — 02H, 02L, 02P, 06I, 06K, 06L, 06Y, 07L, 07Q, 14H, 14J, 19A, 19Y, 22A и 02D. Вместе все ассоциации покрывают 53,5 % транзакций, включающих товар 02D. Средняя точность ассоциаций составляет 0,97. Среднее обеспечение транзакций с товаром 02D — 0,23 (10 % от общего количества транзакций). Корзина № 2 иллюстрируется диаграммой на рис. 4. Корзина № 2 имеет пересечение с корзиной № 1 по товарам 07Q, 06K и 02L.

Поиск следующих корзин осуществляется аналогичным образом.

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Таблица 1. Ассоциации с товаром 01B, выявленные системой Deep Data Diver

№	Associations with 01B	Support		Accuracy
		All rows	Rows with 01B	
1	**01C 01D 01I 02C ⇒ 01B**	0,13	0,38	0,98
2	**01C 01D 01K 02C ⇒ 01B**	0,12	0,35	0,99
3	**01C 01D 01L 02P ⇒ 01B**	0,11	0,31	1
4	**01C 01D 02B 02C ⇒ 01B**	0,12	0,33	0,99
5	**01C 01D 02B 02P ⇒ 01B**	0,11	0,31	1
6	**01C 01D 02C 02L ⇒ 01B**	0,11	0,32	0,99
7	**01C 01D 02C 02P ⇒ 01B**	0,11	0,33	0,99
8	**01C 01D 02L ⇒ 01B**	0,14	0,39	0,98
9	**01C 01I 02C 02P ⇒ 01B**	0,10	0,30	1
10	**01C 01K 02B ⇒ 01B**	0,14	0,41	0,99
11	**01C 01K 06K ⇒ 01B**	0,11	0,31	0,99
12	**01C 01X 02C ⇒ 01B**	0,13	0,38	0,98
13	**01C 02C 02L ⇒ 01B**	0,12	0,34	0,99
14	**01D 01I 02B 02C ⇒ 01B**	0,11	0,30	1
15	**01D 01I 02C 02P ⇒ 01B**	0,10	0,29	1
16	**01D 01K 07A ⇒ 01B**	0,10	0,29	0,99
17	**01D 01X 02L ⇒ 01B**	0,11	0,32	0,99
18	**01I 01K 03X ⇒ 01B**	0,10	0,29	0,99
19	**01I 01L 02B ⇒ 01B**	0,11	0,31	0,98
20	**01I 01X 02C 06K ⇒ 01B**	0,10	0,29	0,99
21	**01K 07Q ⇒ 01B**	0,11	0,32	0,99
22	**02B 06K ⇒ 01B**	0,10	0,29	0,99

4. Выводы

1. Система Deep Data Diver™ использует новую технологию поиска ассоциативных правил, основанную на модифицированном аппарате линейной алгебры с использованием процедуры самоорганизации данных и эффекта информационного структурного резонанса.

2. Уникальные свойства системы позволяют находить в данных высокоточные ассоциации элементов исходного множества транзакций с заданным элементом.

3. Множества ассоциаций с заданными элементами образуют корзины с высоким уровнем обеспечения (support) и длинным набором (long itemsets).

4. На одних и тех же экспериментальных данных показано, что система Deep Data Diver способна выявлять корзины с характеристиками обеспечения и длинами наборов в несколько раз превышающими результаты системы PolyAnalyst.

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

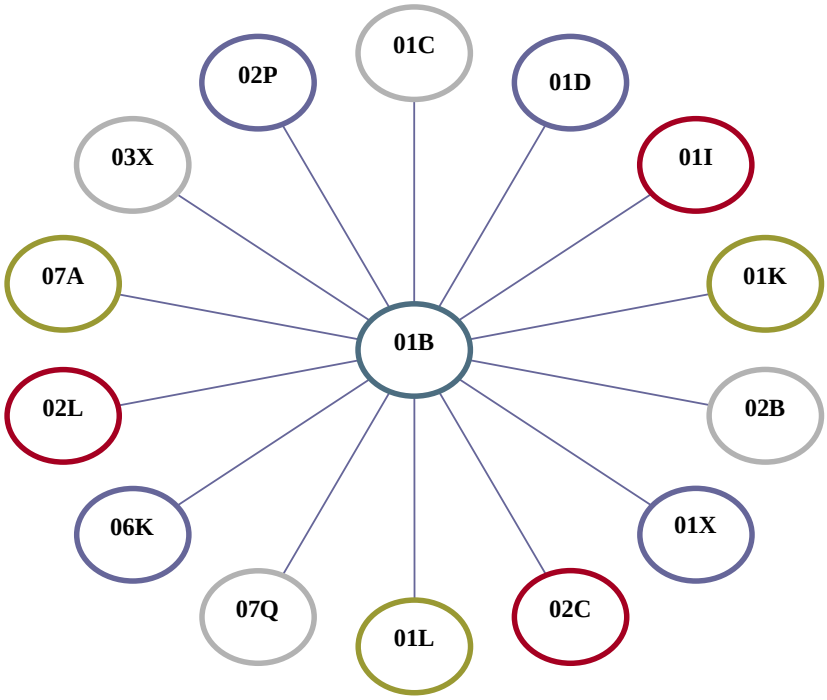


Рисунок 3. Корзина товаров с наиболее часто покупаемым товаром 01B

Таблица 2. Ассоциации с товаром 02B, выявленные системой Deep Data Diver

	Associations	Support 02D	Accuracy
1	02P 06I 06K 06L 06Y ⇒ 02D	0,24	1,0
2	06I 06L 06Y ⇒ 02D	0,32	0,97
3	06L 07L 07Q ⇒ 02D	0,26	0,97
4	02H 06Y 07Q 14H ⇒ 02D	0,20	0,96
5	06K 07Q 19Y ⇒ 02D	0,18	0,98
6	02H 02L 06I 06L 14J ⇒ 02D	0,19	0,96
7	02L 06Y 14J ⇒ 02D	0,26	0,95
8	06I 22A ⇒ 02D	0,22	0,96
9	02L 06Y 19A ⇒ 02D	0,20	0,98
10	06Y 19A ⇒ 02D	0,25	0,95

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

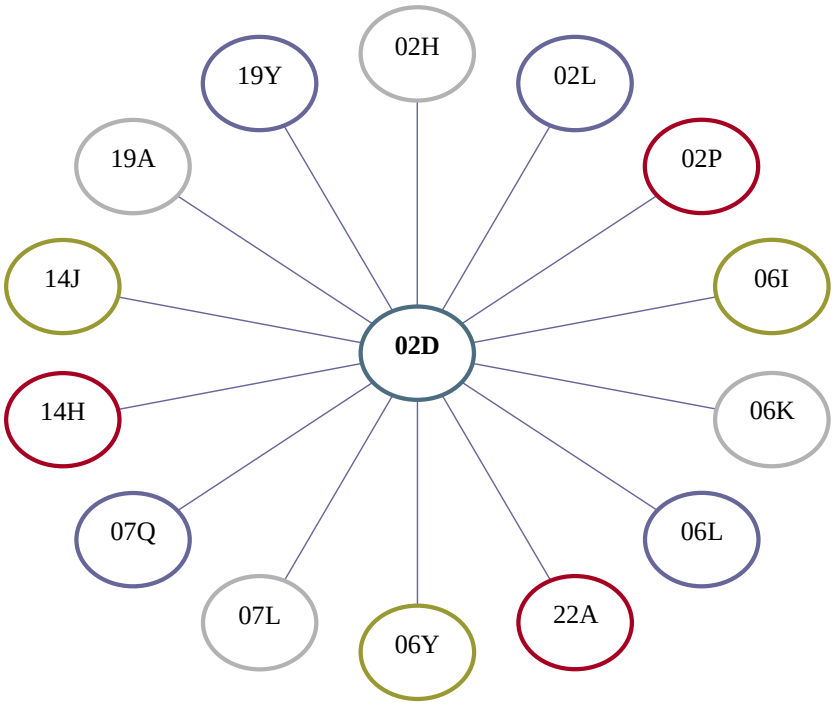


Рисунок 4. Корзина с товаром 02D

Литература

[1] *Ganti V., Gehrke J., Ramakrisnan R.* Mining Very Large Databases // IEEE Computer. August 1999. P. 38–45.

[2] *Charu C. Aggarwal*. Towards Long Pattern Generation in Dense Databases // SIGKDD Explorations. 2001. Vol. 3, Issue 1. P. 20–26,.

[3] *Agraval R., Imielinski T., Swami A.*. Mining Association Rules between Sets of Items in Very Large Databases // ACM SIGMOD Conference Proceedings. 1993. P. 207–216,

[4] *Дюк В. А.* Обработка данных на ПК в примерах. СПб.: Питер, 1997. 240 с.

[5] *Duke V. A.* Latent knowledge extraction by methods of local geometry: development of expert system for keen appendicitis diagnostics // Proc. Intern. Conf. On Informatics and Control (ICI&C 97), 1997. St. Petersburg, Russia. Vol.2, P. 663–668.

[6] PolyAnalyst Tutorials. — Megaputer Intelligence, <http://www.megaputer.com>.