УДК 004.6

А. П. Кирпичников, И. С. Ризаев, З. Т. Яхина,
А. Л. Осипова

ИССЛЕДОВАНИЕ ЗАКОНОМЕРНОСТЕЙ МЕЖДУ СВЯЗАННЫМИ СОБЫТИЯМИ

Ключевые слова: 1,2,3-ассоциативные правила, поддержка, достоверность.

Рассматриваются ассоциативные правила, которые позволяют находить закономерности между связанными событиями или объектами.

Keywords: 1,2,3-Association rules, support, reliability

Discusses the Association rules, which allow us to find patterns between related events or objects.

Введение

При обработке информации в больших базах данных возникает задача исследования закономерностей между связанными событиями. Современные баз данных могут иметь весьма внушительные объемы. Одним из популярных методов обнаружения знаний стали алгоритмы поиска ассоциативных правил. Первоначально задача поиска ассоциативных правил была предложена для нахождения типичных шаблонов покупок в супермаркетах, что было объединено в наименование *анализ рыночной корзины* (market basket analysis). Например, покупатель купивший молоко, как правило, покупает еще кефир и (или) творог. Примером такого правила, служит утверждение, что покупатель, приобретающий «Хлеб», приобретет и «Молоко» с вероятностью 75%. Поскольку реальные базы данных транзакций, рассматриваемые при анализе рыночной корзины, обычно содержат тысячи предметов, вычислительные затраты при поиске ассоциативных правил огромны. Поэтому является важным в процессе генерации ассоциативных правил использовать методики, позволяющие уменьшить количество ассоциаций, которое требуется проанализировать.

Поиск ассоциативных правил

Задачей поиска ассоциативных правил является нахождение закономерностей между связанными событиями или объектами [1].

Примерами приложения ассоциативных правил могут быть следующие задачи:

- розничная торговля: покупка совместных товаров; анализ потребительской корзины; прогнозирование спроса;

- классификация групп людей: выявление групп покупателей; выявление общих характеристик людей собранных для выполнения сложной проблемы в компании; выявление характеристик людей склонных к совершению противоправные действия и т.д.

- набор лекарственных средств и процедур, используемых больными при лечении определенных заболеваний;

- заключение о наличии или отсутствии нефти при исследовании площади по взятым пробам из многочисленных скважин;

- определение профиля посетителей веб-ресурсов и т.д.

Базовые понятия

Базовым понятием в теории ассоциативных правил является – *транзакция*, представляющая собой множество событий совершаемых одновременно [2]. Типичным примером является покупка товаров, обычно покупка некоторого множества товаров в супермаркете, называемой *рыночной корзиной*.

Так в табл.1 приведен пример «рыночной корзины». На основе имеющейся базы данных нужно найти закономерности между событиями, связанными с покупками товаров.

Таблица 1 – Пример набора транзакций

№ транзакции	Транзакции (покупки)
101	Молоко, кефир, творог, сметана
102	Батон, печенье, конфеты
103	Молоко, творог
104	Огурцы, помидоры, салат
105	Молоко, сметана, творог
106	Помидоры, огурцы
107	Молоко
107	Помидоры, огурцы, лук
…	…
200	Молоко, творог, сыр

Визуальный анализ показывает, что некоторые продукты, как правило, покупаются совместно. Если покупаются огурцы, то покупаются и помидоры. Ассоциативное правило записывается в виде условия X → Y «если X то Y», что формулируется в виде: «Если условие, то следствие». Ассоциативное правило описывает связь между предметами, соответствующими условию и следствию. Например «огурцы → помидоры», «молоко →сметана, творог».

Введем формальное описание задачи поиска ассоциативных правил:

Пусть A = {a_1, a_2, … ,a_j, … , a_n} множество исследуемых объектов;

T = {t_1, t_2, … ,t_i, … , t_n} – множество транзакций; $\forall a_j \in t_i \Rightarrow a_j \in I$,

где I – все множество объектов, входящих в базу данных.

Для оценки полезности ассоциативных правил вводятся следующие величины:

1. *Поддержка (support)* – показывает, какой процент транзакций поддерживает данное правило:

$$Supp_{X\Rightarrow Y} = \frac{T_{X\cup Y}}{T}$$

2. *Достоверность (confidence)* – показывает, какова вероятность того, что из события *X* следует событие *Y*:

$$Conf_{X\Rightarrow Y} = \frac{T_{X\cup Y}}{T_X} = \frac{Supp_{X\cup Y}}{Supp_X}$$

3. *Улучшение (improvement)* – показывает насколько полезнее полученное правило по сравнению со случайным угадыванием:

$$\mathrm{Im}\, pr_{X\Rightarrow Y} = \frac{T_{X\cup Y}}{T_X \cdot T_y} = \frac{Supp_{X\cup Y}}{Supp_X \cdot Supp_Y}$$

Анализ транзакций, представленных в табл.1, выявил с достоверностью 50% два правила:

1. Молоко →творог (3 транзакции с поддержкой 15%).

2. Огурцы →помидоры (1 транзакция с поддержкой 5%).

Для поиска ассоциативных правил был применен пакет Deductor Academic 5.3 (www.basegroup.ru).

Пусть имеется набор объектов сгруппированных в транзакции (табл. 2).

Таблица 2 – Набор транзакций

№ транзакции	Транзакции
1	a, c
2	b, d, c
3	b, a, d, c
4	a, c, d
5	d, c

Набор {a, c} встречается в транзакциях 1, 3, 4 всего три раза, что определит поддержку величиной $Supp_{(a->c)}$= 3/5. Поддержка набора {b, d} составит $Supp_{(b->d)}$= 2/5. Поддержка набора {a, c, d} составит $Supp_{(a->c,d)}$= 2/5.

Аналогично $Supp_{(a)}$=3/5; $Supp_{(c)}$=5/5=1.

Аналитик при решении задачи поиска может указать минимальное значение поддержки – SuppMIN и рассматривать только наборы для которых выполняется условие: Supp(T)>SuppMin.

Различные наборы, встречающиеся одинаковое число раз, могут иметь одинаковое значение поддержки. Для повышения значения правила используется – *достоверность*.

Так достоверность наборов составит:
$Conf_{(a->c)}$ =1; $Conf_{(a->c,d)}$=2/3; $Conf_{(c->a,d)}$=2/5.

Считается, что чем выше достоверность, тем лучше правило. Рассмотрим impr (улучшение) для набора {a, c}.

$$impr_{a\to c} = \frac{Supp_{a\cup c}}{Supp_a \cdot Supp_c} = \frac{3/5}{3/5\cdot 1} = 1$$

Полученное значение указывает, что если impr≥1, то с помощью правила предсказать наличие набора "Y" после "X" вероятнее, чем случайное угадывание, если меньше единицы, то наоборот.

Задачами нахождения ассоциативных правил являются: поиск всех возможных групп наборов и составление правил для найденных групп [2].

Надо иметь в виду, что число возможных ассоциаций с увеличением числа предметов быстро растет экспоненциально. Количество групп G из множества N объектов можно определить по формуле:

$$G = \sum_{d=1}^{N}\binom{N}{d} = \sum_{d=1}^{N} {}^{d}_{N} = \sum_{d=1}^{N}\frac{N!}{d!(N-d)!}$$

Количество возможных правил R из множества N объектов находят по формуле:

$$R = \sum_{d=1}^{N}\binom{N}{d}\cdot\sum_{d=1}^{N} {}^{d}_{N} = \sum_{j=1}^{N-d}\binom{N-d}{j}$$

Более просто количество возможных правил можно определить по формуле

$$R=3^{N}-2^{N+1}+1.$$

Например, при *N*=10, количество правил составит *R*=57002. Так при минимальной поддержки в 20% и минимальной достоверности 50%, оказывается, что 80% этих правил не являются полезными. Поэтому при поиске правил необходимо стремиться к уменьшению количества групп.

Существуют алгоритмы, которые позволяют найти последовательности, удовлетворяющие условию отбора больше чем SuppMIN, например AprioriALL и AprioriSome [3].

Таблица 3 – Представление частых элементов

№ транзакции	Транзакции	Представление частых элементов			
1	a, c	c		a	
2	b, d, c	c	d		b
3	b, a, d, c	c	d	a	b
4	a, c, d	c	d	a	
5	d, c	c	d		
		5	4	3	2

Возьмем и перестроим таблицу 2 по частоте элементов таблицы 3. На рис.1. изобразим последовательности в виде графа.

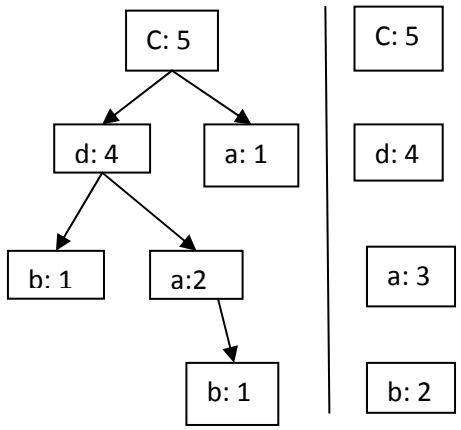

Рис. 1 – Последовательности наборов

Если использовать минимальный уровень величины транзакции равной трем объектам, то все наборы меньше этой величины можно отбросить. Так в данном случае (рис.3) сюда попадут последовательности включающие элемент «b», это {b, d, c} и {b, a, d, c}.

Можно указать следующий алгоритм поиска часто встречающихся наборов объектов [2].

1. Сканировать все транзакции и произвести отбор всех 1-элементных транзакций, у которых поддержка больше минимально заданной величины (k=1 – количество объектов в группе).

2. Исключить элементы, которые не отвечают требованиям поддержки.

3. Задать k=k+1. Найти все возможные транзакции удовлетворяющие условию поддержки.

4. Повторять этапы 2 и 3 до тех пор, пока не дойдем до самой длинной транзакции.

Выводы

Современные базы данных могут содержать огромное количество накопленной информации. При правильном подходе можно получать не только оперативную информацию, но и извлекать новые знания (Data Mining) [4]. Одним из таких подходов является поиск закономерностей между связанными событиями или объектами. Ассоциация является одной из задач Data Mining, предназначенной для извлечения ценной информации в базе данных. Найти закономерности между событиями или объектами методом перебора может завести вычислительные ресурсы в тупик. Для решения таких задач и предлагается подход, связанный с поиском ассоциативных правил на основе поддержек и достоверностей.

Литература

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод. И.И. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP/ Изд-во СПБ.: БХВ-Петербург. – 2008. – 384 с.
2. Ризаев И.С., Рахал Я. Интеллектуальный анализ данных для поддержки принятия решений./Казань: Изд-во МОиН РТ. 2011. – 172 с.
3. Agrawal, Srikant R. Fast Algorithma for Mining Association Rules in large databases// 20th Int'l Conf. of Very Large Data Bases, Sept. – 1994.
4. Кирпичников А.П., Ризаев И.С., Осипова А.Л.. Повышение аналитических возможностей баз данных / Вестник Казан. технол. ун-та. 2012. **15**, –С.157-160.

© **А. П. Кирпичников** – д-р физ.-мат. наук, проф., зав. каф. ИСУИР КНИТУ, kirpichnikov@kstu.ru; **И. С. Ризаев** – к.т.н., проф. каф. АСОИУ КНИТУ-КАИ, isr4110@mail.ru; **З. Т. Яхина** – к.т.н., доц. той же кафедры; **А. Л. Осипова** – ст. преп. той же кафедры.

© **A. P. Kirpichnikov** - Dr. Sci, Head of the Department of Intelligent Systems & Information Systems Control, KNRTU, kirpichnikov@kstu.ru; **I. S. Rizaev** – PhD, Professor of the Department of Automated Information Processing Systems & Control, KNRTU named after A.N. Tupolev, isr4110@mail.ru; **Z. T. Yahina** - PhD, Associate Professor of the Department of Automated Information Processing Systems & Control, KNRTU named after A.N. Tupolev; **A. L. Osipova** – Senior Lecturer of the Department of Automated Information Processing Systems & Control, KNRTU named after A.N. Tupolev.