УДК 681.3.06

Поиск ассоциативных правил при анализе баз данных

Ильенков В.И, Ковтун М.А., Придухо В.Т.
Белорусский национальный технический университет

Поиск ассоциативных правил является одним из методов решения задачи классификации с целью определения часто встречающихся наборов объектов при наличии существенного множества таких наборов. При анализе этих данных важно получить информацию о том, какие объекты взаимодействуют вместе, в какие периоды времени и в какой последовательности. Наиболее известным алгоритмом для решения этих задач является алгоритм Apriori. В тоже время реализация алгоритма должна предусматривать хранение и анализ данных из разных источников – текстового, xml, MSSQl Server и др. Исходя из этого для поиска правил выбираем для реализации алгоритм Apriori с проведением его модификации, чтобы реализовать возможность анализа данных из разных источников.

На первом шаге алгоритма подсчитываются одноэлементные часто встречающиеся наборы. Для этого необходимо пройтись по всему набору данных и подсчитать для них поддержку, т.е. сколько раз набор встречается в базе. Следующие шаги состоят из двух частей: генерации потенциально часто встречающихся наборов элементов (их называют кандидатами) и подсчета поддержки для кандидатов.

Приложение было протестировано на примере анализа базы данных магазина подарков, имеющей более 15000 транзакций. Выявленные закономерности представлялись в виде правил «если,... то». Для каждого правила определялись и отображались такие параметры, как поддержка и достоверность. В докладе проведено сравнение разработанной программы и MS SQL Server – реализацией от Microsoft, который содержит набор служб Analysis Service, реализующие различные алгоритмы DataMining, в том числе и поиск ассоциативных правил.

Табличная визуализация от Microsoft, практически аналогична визуализации в разработанном приложении, результаты почти полностью совпадают.

Визуализация от MS SQL Server выглядит красиво, но она менее информативна, чем представление в виде дерева в приложении. В нем в дополнение к интуитивно понятному представлению, присутствуют еще и цифры (достоверность, поддержка), позволяющие легко сориентироваться и принять правильное решение, без дополнительных просмотров табличных данных.