ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕТОДИК ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ХРАНИЛИЩ ДАННЫХ, СОДЕРЖАЩИХ НЕСКОЛЬКО ТАБЛИЦ ФАКТОВ

Отчет по проекту N 16-37-50067

Михаилов М.В., Чечулин А.А.

2016 г.

Введение

В настоящее время информационно-аналитические системы являются важной и неотъемлемой частью практически всех коммерческих и государственных процессов. При этом важным фактором является стремительный рост объема и усложнение структуры хранимых данных. Для обработки запросов к таким хранилищам часто используют OLAP (англ. online analytical processing, аналитическая обработка в реальном времени), который представляет собой технологию обработки данных, заключающаяся в подготовке суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу. OLAP-структура, формируемая на основе рабочих данных, называется OLAP-куб. Куб создается из объединения таблиц в виде таких схем представления многомерных данных, как звезда, снежинка и созвездие. В первой и второй схемах создается большая центральная таблица, называемая таблица фактов (fact table). В нее записываются основные данные, к которым будут выполняться пользовательские запросы. Для этих схем существует множество эффективных методик анализа данных, и эти методики широко представлены в существующих публикациях. При этом методы интеллектуальная обработка хранилища данных со структурой «созвездие» сравнительно мало изучена и на данный момент не хватает методик анализа таких типов хранилищ, а изучить связи между таблицами фактов зачастую

очень важно для выявления скрытой информации о процессе, который описан изучаемым хранилищем данных.

Таким образом, одной из фундаментальных научных проблем является отсутствие эффективных методик интеллектуального анализа хранилища данных со структурой «созвездие» (fact constellation schema), т.е. содержащих несколько таблиц фактов. Изучение данной проблемы связано еще и со сложностью выгрузки данных из хранилища данных (Data Warehouse) в системы интеллектуального анализа (Vincent Rainardi «Building a Warehouse», Apress, 2008), а также хранением агрегатов для последующего анализа. Кроме того, очень важной задачей является поиск скрытых связей между таблицами фактов в хранилище. Эти связи могут быть представлены в виде ассоциативных правил, сложных последовательностей и кластеров. При делении на кластеры, в настоящем исследовании, предложено использование мер близости, которые основаны на коэффициенте корреляции, а также вейвлет-коэффициентах. Это позволяет находить связи, которые имеют запаздывание по времени относительно друг друга. Для этой цели в работе предложена разработка кластеризационно-корреляционного сканера хранилища данных. Отдельного внимания заслуживает поиск ассоциативных правил объединяющих записи различных таблиц. Такой поиск построен на расширении существующих алгоритмов, основанных на переборе данных. Также в рамках данного исследования был рассмотрен вопрос применения различных эвристических подходов к поиску ассоциативных правил в хранилищах данных, содержащих несколько таблиц фактов, для сокращения количества перебора.

1. Методика выбора сочетания реляционного и многомерного хранения данных

Одной из проблем интеллектуального анализа является сложность выгрузки данных из OLAP-хранилищ в системы интеллектуального анализа (BI-системы), которые имеют следующие особенности [1]:

- необходимость работы в режиме реального времени в рамках процесса приема-передачи данных;
- необходимость быстрого выполнения аналитических запросов;
- необходимость возможности поиска скрытых закономерностей среди элементов хранилища.

С учетом перечисленных особенностей, сложность заключается в скорости работы, которая существенно снижается из-за переполнения хранилища агрегатными значениями (“взрыв” хранилища), или неоптимальной структуры хранилища. Таким образом, обеспечение высокой скорости выгрузки сводится к построению методики организации хранения данных на основе выбора структуры хранения данных и выбора оптимального способа физического хранения данных.

Рассмотрим способы физического хранения данных. Существует три основные концепции организации OLAP хранилища [2]:

- Relational OLAP (ROLAP) – организация хранения OLAP-структуры многомерного куба данных в реляционной БД;
- Multidemsional OLAP (MOLAP) – организация хранения OLAP-структуры многомерного куба данных в многомерной БД;
- Hybrid OLAP (HOLAP) – организация хранения OLAP-структуры многомерного куба данных в гибридной БД.

Рассмотрим их достоинства и недостатки. При этом, критериями обеспечения высокой скорости выгрузки будут являться защищенность от “взрыва” данных, а также высокая скорость работы с агрегированными и не агрегированными данными.

Достоинством ROLAP-систем является хранение данных в реляционных таблицах, что позволяет использовать уже существующие БД компаний. Как правило, для ускорения агрегации, ROLAP-хранилища содержат множество дополнительных таблиц с некоторыми заранее агрегированными наборами данных. С одной стороны, преимущество данного подхода состоит в том, что он позволяет работать с большими объемами данных. С другой стороны, подход имеет существенный недостаток – количество вспомогательных таблиц многократно превышает количество таблиц с данными (т.е. приводит к “взрыву” хранилищ). Так, использование ROLAP приводит к увеличению объема хранилищ от 300 до 1200% [3]. Также, недостатком является то, что из-за трансляции MDX-запросов в SQL-запросы, функциональность ROLAP-системы ограничивается возможностями языка SQL.

Системы Multidemsional OLAP (MOLAP) имеют другую структуру. В них многомерный куб хранится непосредственно в многомерной БД. Таким образом, преимуществом данной системы является то, что MDX-запросы выполняются над многомерным кубом, а не реляционной БД, что позволяет существенно увеличить скорость запросов к данным. Системы MOLAP, аналогично ROLAP, хранят как агрегированные данные, так и не агрегированные. Так как элементы MOLAP являются копиями элементов реляционной БД, она требует дополнительного дискового пространства. Однако, объем используемого дискового пространства у MOLAP на порядок меньше, чем у ROLAP, что также является преимуществом. С другой стороны, использование MOLAP подразумевает неизменность существующих данных: данные не могут быть изменены, только добавлены; что не всегда достижимо. В случае изменения данных, многомерный куб необходимо полностью перестроить, что является существенным недостатком. Еще одним недостатком является то, что так как MOLAP собирает информацию из реляционных БД, в отличие от ROLAP, для его построения необходимы дополнительные инструменты.

Тип Hybrid OLAP (HOLAP) наследует преимущества ROLAP и MOLAP. HOLAP использует сразу оба типа БД: многомерную БД для агрегированных данных и реляционную БД для не агрегированных. HOLAP имеет явное преимущество, которое выражается в том, что HOLAP позволяет исключить копирование не агрегированных данных из реляционной БД в многомерный куб. Таким образом, повышается скорость доступа к агрегированным данным по сравнению с MOLAP, однако скорость доступа к не агрегированным данным снижается, что является недостатком. Стоит отметить, что гибридное использование ROLAP и MOLAP подразумевает их взаимодействие, что повышает сложность реализации и также является недостатком.

Таким образом, с точки зрения выгрузки данных в системы интеллектуального анализа оптимальным будет использование гибридного хранилища (HOLAP). Использование одновременно и реляционного и многомерного типов хранилищ позволяет с высокой скоростью получать доступ к агрегированным значениям, и с достаточной скоростью к не агрегированным. Также использование HOLAP позволяет избежать “взрыва” хранилища, так как хранилище не переполняется агрегированными значениями.

Однако одного использования HOLAP-хранилища недостаточно для обеспечения высокой скорости выгрузки. Немаловажным критерием является построение оптимальной структуры данных, использование которой было бы эффективно при обработке, выявлении и создании необходимой информации в процессе интеллектуального анализа.

Сам интеллектуальный анализ базируется на комбинации и определённой последовательности применения следующих методов:

- ассоциации – сопоставления двух элементов схожего типа;
- классификации – определение типа элемента;
- кластеризации – группировка элементов;
- прогнозирования – анализ тенденций.

Таким образом, для обеспечения высокой скорости выгрузки данных из OLAP-хранилищ в BI-системы была разработана методика выбора сочетания реляционного и многомерного хранения данных основанная на использовании реляционного и многомерного хранения данных (HOLAP-хранилищах) и создания соответствующей оптимальной структуры данных.

Для методов интеллектуального анализа оптимальным будет представление элементов в HOLAP-хранилище в виде таблиц имеющих набор числовых и нечисловых параметров. Выбор системы параметров не автоматизирован и производится человеком исходя из бизнес-требований и требований к данным, хотя конечно их значения могут вычисляться автоматически. Логические процессы методики выбора сочетания реляционного и многомерного хранения данных отображены на рисунке 1.

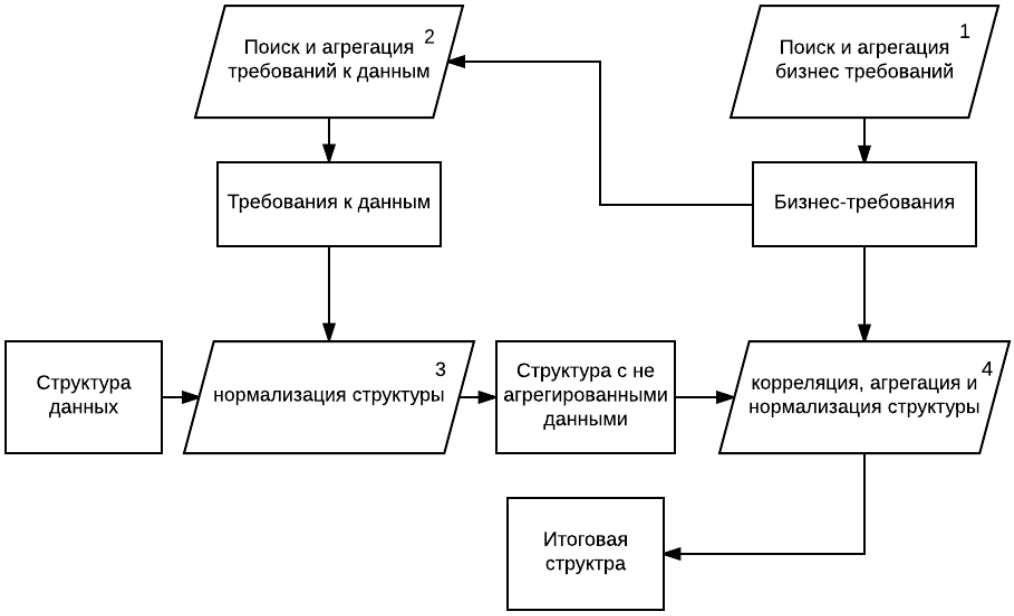

Рисунок 1 – методика выбора сочетания реляционного и многомерного хранения данных

Основные этапы методики:

1. агрегация бизнес-требований;
2. агрегация требований к данным;
3. нормализация структуры данных в соответствии с требованиями к данным;

4. корреляция, агрегация и нормализация структуры данных в соответствии с требованиями бизнес требований.

Этап 1.

Бизнес-требования заключаются в поиске и создании требований к новым переменным, которые необходимы для анализа. Одна задача анализа представляет собой цепочку аналитических операций (ассоциации, классификации, кластеризации и прогнозирования). Каждый элемент цепи требует определённого набора аналитических переменных, которые можно агрегировать. Прежде чем составлять набор этих переменных необходимо обозначить переменные какого типа нужны, необходимо ли создавать новые переменные и можно ли агрегировать существующие. Таким образом, первый этап методики сводится к созданию требований к срезам агрегированных данных.

Этап 2.

Второй этап методики сводится к созданию требований к срезам и полям не агрегированных данных, на основе которых можно получить агрегированные данные. Требования к данным также должны учитывать бизнес-требования. В свою очередь не агрегированные данные должны храниться в реляционной составляющей, а агрегированные (например, срезы и свойства наборов данных) в многомерной составляющей HOLAP-хранилища.

Этап 3.

На третьем этапе методики происходит нормализация структуры данных в соответствии с требованиями к данным. Фактически происходит формирование структуры таблицы в соответствии с требованиями, которые были установлены на предыдущем этапе. В результате должна получиться таблица не агрегированных данных, структура которой формально выглядит следующим образом:

$$P_i = \{p_{i,1}, p_{i,2}, p_{i,3}, \dots, p_{i,j_i}, \dots, p_{i,n_i}\}, i = \overline{1,m}, j_i = \overline{1,n_i} \quad (1)$$

, где P_i - это набор полей таблицы состоящий из m полей, а p_i, j_i - одно из n_i количества значений для поля P_i.

Этап 4.

На четвертом этапе методики происходит корреляция, агрегация и нормализация структуры данных в соответствии с бизнес требованиями которые установлены на первом этапе. Структура дополняется агрегированными полями полученных за счет объединения записей различных таблиц на основе ассоциативных правил. При этом могут быть добавлены поля содержащие информацию о принадлежности к кластеру. Итоговая структура не должна иметь поля которые не применяются в процессе анализа, а также структура должны содержать объединенные наборы, которые удовлетворяют бизнес-требованиям. Наборы формально выглядят следующим образом:

$$T_k = \{p_i, j_i\}, k = \overline{1,K}, i = \overline{1,m} \quad (2)$$

, где K - количество наборов, k - номер набора, а j_i - номер значения из списка допустимых значений для таблицы P_i.

Таким образом, структура данных, полученная в результате работы методики, в совокупности с использованием HOLAP-хранилища позволяет ускорить процесс выгрузки с системы интеллектуального анализа.

2. Методика кластеризации временных рядов, полученных из хранилища данных, по мерам близости, основанным на коэффициенте корреляции и вейвлет-коэффициентах

Основная сложность кластеризации временных рядов заключается в выборе меры близости между данными и в том, что кластеризация должна происходить не на основе элементов временного ряда, а на основе некоторых его характеристик, которые отображают тенденцию. При этом немаловажным является возможность характеристик противостоять влиянию выбросов и различию в масштабах рядов относительно тренда. Для решения данной проблемы была разработана методика кластеризации временных рядов, полученных из хранилища данных, по мерам близости. Методика принимает на вход меры близости, которые основаны либо на коэффициенте корреляции, либо на вейвлет-коэффициентах.

Рассмотрим алгоритм кластеризации временных рядов, полученных из хранилища данных по мерам близости.

Пусть характеристики временного ряда даны в виде вектора коэффициентов $\widetilde{W}_{j,k}^{R_i}$, где R_i – ряд к которому относятся коэффициенты, j и k – значения масштаба временного ряда. Таким образом, мера близости будет определяться на основе векторов коэффициентов между двумя рядами, которые имеют одинаковый масштаб. Для ее вычисления будет использоваться косинусное сходство. Таким образом, сходство между векторами A и B определяется по формуле использующей косинусное произведение и норму:

$$\cos(\theta) = \frac{\widetilde{W}_{j,k}^{A} \cdot \widetilde{W}_{j,k}^{\mathrm{B}}}{\left\|\widetilde{W}_{j,k}^{A}\right\| * \left\|\widetilde{W}_{j,k}^{B}\right\|} = \frac{\sum_{i=1}^{n} \widetilde{W}_{i}^{A} * \widetilde{W}_{i}^{B}}{\sqrt{\sum_{i=1}^{n}\left(\widetilde{W}_{i}^{A}\right)^2} * \sqrt{\sum_{i=1}^{n}\left(\widetilde{W}_{i}^{B}\right)^2}} \quad (3)$$

Мера близости определяется как значение, для которого большее значение свидетельствует о дальности объектов и вычисляется по формуле:

$$D = -\ln(\cos\theta) \tag{4}$$

Таким образом можно судить и сходстве или различии временных рядов исходя их характеристик. Рассмотрим методики их получения.

Рассмотрим алгоритм получения характеристик временного ряда на основе вейвлет-коэффициентов [4].

Пусть дан R – временной ряд длины $\mathrm{N} = 2^{\mathrm{m}}$. При этом, понятно, что не для любого N можно подобрать целочисленное m. Поэтому на первом шаге ряд R дополняется нулями или значениями, которые равны среднему $\overline{\mathrm{X}}$ ряда

$$\overline{\mathrm{X}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=0}^{\mathrm{N}-1}\mathrm{x_n} \tag{5}$$

, до тех пор пока не будет выполняться условие $\mathrm{N} = 2^{\mathrm{m}}$.

На следующем шаге с использованием фильтра Дебеши подсчитываются вейвлет-коэффициенты. Вейвлет фильтр длины L формально выглядит следующим образом:

$$\{\mathrm{h_n}: \mathrm{n} = 0, \dots, \mathrm{L}-1\} \tag{6}$$

Масштабный фильтр длины L формально выглядит следующим образом:

$$\{\mathrm{g_n} \equiv (-1)^{\mathrm{n}+1} * \mathrm{h_{L-n-1}}, \mathrm{n} = 0, \dots, \mathrm{L}-1\} \tag{7}$$

Вычисление вейвлет-коэффициентов производится по формулам дискретного вейвлет преобразования:

$$W_{j,k}^{R} = \sum_{l=0}^{L-1} h_l * V_{j-1,(2*k+1-l)\mathrm{mod}N_{j-1}} \quad (8)$$

где,

$$V_{j,k} = \sum_{l=0}^{L-1} g_l * V_{j-1,(2*k+1-l)\mathrm{mod}N_{j-1}} \quad (9)$$

Далее исключаются граничные вейвлет коэффициенты, а также коэффициенты, при подсчете которых использовались нули или средние значения временного ряда. Таким образом, происходит получение временных характеристик, основанных на вейвлет коэффициентах, необходимых для кластеризации временных рядов.

Рассмотрим алгоритм получения характеристик временного ряда на основе коэффициентов корреляции.

Предлагается для вычисления вектора коэффициентов использовать корреляцию Пирсона между последовательными элементами ряда. Пусть дан R – временной ряд длины N состоящий из элементов r_i. Корреляция будет вычисляться между последовательными наборами элементов P длиной n. Таким образом, вычисление будет производиться по формуле:

$$W_i^R = \frac{cov(x,y)}{\sqrt{s_x^2 * s_y^2}} \quad (10)$$

где x и y последовательные непересекающиеся наборы длины n:

$$x = (r_i, \dots, r_n), y = (r_{n+1}, \dots, r_{n*2}) \quad (11)$$

s_x^2 и s_y^2 - выборочные дисперсии в которых используются выборочные средние $\bar{x}$ и $\bar{y}$ соответственно:

$$s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2, s_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad (12)$$

а $cov(x, y)$ является ковариацией между наборами:

$$cov(x, y) = \sum_{i=0}^{n}(x_i - \bar{x}) * (y_i - \bar{y}) \quad (13)$$

Итоговое вычисление сводится к формуле:

$$W_i^R = \frac{\sum_{i=0}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (14)$$

Таким образом, происходит получение временных характеристик, основанных на коэффициентах корреляции, необходимых для кластеризации временных рядов.

Для сравнения полученных мер близости были проведены эксперименты, где в качестве исходных данных выступали котировки по основным валютам за период с 01.01.2016 по 01.08.2016, взятые с сайта www.finam.ru.

Сравнение мер показало, что меры близости, основанные на корреляции, лучше работают на небольших отрезках и отображают сходство как последовательное сходство локальных трендов. При этом мера позволяет производить классификацию на волатильных валютах.

Меры близости, основанные на вейвлет коэффицентах, также показали наличие робастности, и, в отличие от мер основанных на корреляции, учитывают схожесть локальных трендов с глобальным трендом.

Таким образом, в зависимости от типа задач предлагается в кластеризационно-корреляционном сканере использовать меры близости, основанные на вейвлет коэффициентах, а также на коэффицентах-корреляции.

Для разработки кластеризационно-корреляционного сканера на мерах близости, необходимо провести сравнительный анализ существующих алгоритмов кластеризации и классификации.

Для реализации кластеризационно-корреляционного сканера были рассмотрены следующие алгоритмы классификации:

- Decision Trees (DT) – классификация объектов путем построения деревьев решений;
- Naive Bayes (NB) – применение вероятностных подходов для классификации объектов;
- Adaptive Bayes Networks (ABN) – расширенный вариант алгоритма NB;
- Support Vector Machines (SVM) – метод опорных векторов;
- Enhanced k-Means Clustering – кластеризация на основе алгоритма k-Means для обнаружения групп схожих объектов;
- Orthogonal Partitioning Clustering – кластеризация методом ортогонального деления;
- Anomaly Detection – анализ исключений (аномальных событий).

Алгоритмы Decision Trees, Naive Bayes, Adaptive Bayes Networks и Support Vector Machines относятся к классу методов, которые обучаются “с учителем”, и требуют некоторую начальную выборку для обучения. Таким образом, они не подходят для задач кластеризации данных в OLAP-структурах.

Алгоритмы Enhanced k-Means Clustering и Orthogonal Partitioning Clustering не требуют обучающей выборки, однако им необходимо изначально указывать на какое количество кластеров нужно разбить данные, что не всегда бывает известно. Однако, методы кластеризации все же лучше подходят для анализа OLAP-структур, чем методы классификации.

Анализ исключений относится к другому классу методов, и может применяться для выделения зависимостей в аномалиях, которые были обнаружены при кластеризации данных.

Алгоритмы кластеризации можно разделить на следующие группы:

1. иерархические:
    a. агломеративные методы;
    b. девизимные методы;
2. неиерархические;
3. адаптивные.

Иерархические методы позволяют выявить вложенные кластеры, совокупность которых можно представить в виде дерева и, следовательно, описать OLAP-структуру в виде типа “звезда” или “снежинка”, которые также имеют древовидное построение. За счет поэтапной кластеризации, иерархические методы позволяют заметно улучшить скорость работы на больших и не численных выборках. Поэтому, как правило, в системах анализа данных используются именно иерархические методы кластеризации.

Неиерархические методы позволяют представить структуру в виде отдельных непересекающихся кластеров, однако скорость их работы над многомерными, большими и не численными агрегатами (которые часто встречаются в OLAP-структурах) неудовлетворительна. Их можно использовать над небольшими срезами, однако для анализа больших многомерных структур они не применимы.

Адаптивные методы позволяют выбрать в зависимости от характеристик данных какой подход (иерархических или неиерархический) следует выбрать. Они позволяют разбить структуру на иерархическую систему

кластеров, а после выполнять неиерархическую кластеризацию над выборками меньшего размера (листьями дерева кластеров). С точки зрения анализа OLAP-структур типа “созвездие” лучше использовать адаптивные методы кластеризации, так как они позволяют описать OLAP-структуру в том же виде («созвездие»). При этом адаптивные методы кластеризации обеспечивают достаточную для интеллектуального анализа скорость работы.

Таким образом, в качестве кластеризационно-корреляционного сканера предлагается методика кластеризации, которая дополняет методику адаптивной кластеризации Adakl [5]. В частности предлагается для расчета матрицы взаимных расстояний использовать алгоритм косинусного сходства, которому на вход подаются вектора мер близости, вычисленные из временных рядов по вышестоящим алгоритмам.

Общая схема методики представлена на рисунке 2.

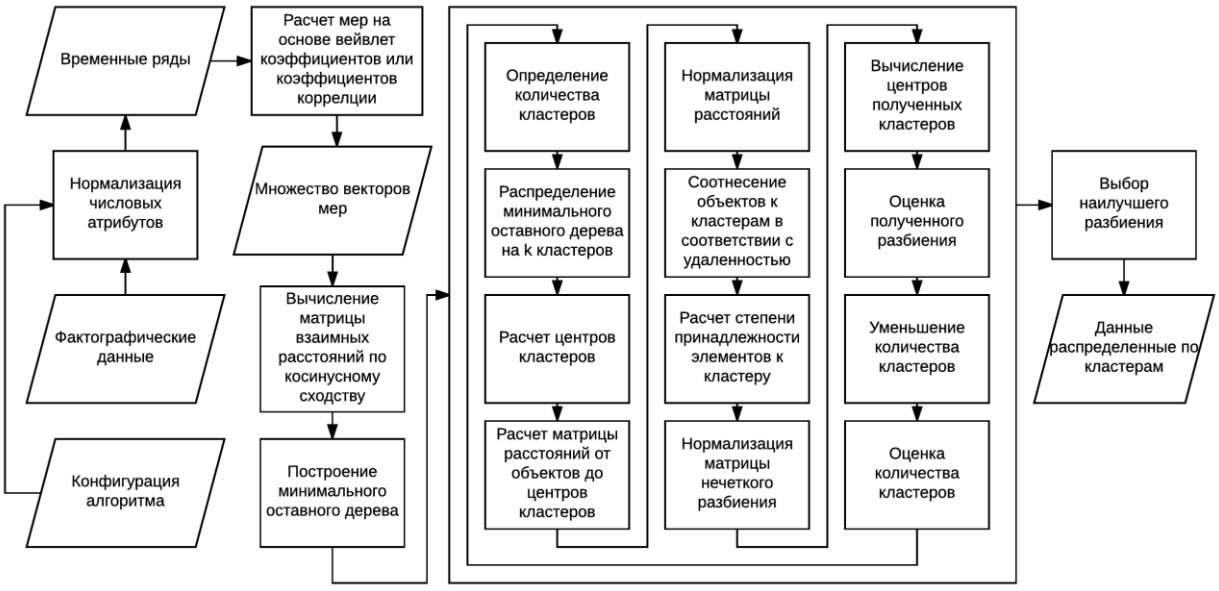

Рисунок 2 - Методика адаптивной кластеризации Adakl [5]

Основные шаги предлагаемой методики:

1) нормализация полученных данных и конфигурация алгоритма – на данном этапе выбираются используемые меры, а также производится преобразование данных к единому масштабу;

2) расчет мер на основе вейвлет-коэффицентов или коэффициентов корреляции – на данном этапе для каждого временного ряда строится вектор характеристик, на основе которых будет происходить расчет расстояний;
3) вычисление матрицы взаимных расстояний по косинусному сходству – расчет расстояний (сходства) между временными рядами, на основе которых будет происходить кластеризация;
4) построение минимального оставного дерева – построение оставного дерева, в котором весу ребра соответствует косинусное сходство;
5) применение методики адаптивной кластеризации Adakl – нахождение оптимального разбиения на кластеры.

Таким образом, кластеризационно-корреляционный сканер на основе разработанной методики позволяет представлять результаты в виде пригодном для хранения в OLAP-структурах типа созвездие, динамически находить требуемое количество кластеров, производить анализ за приемлемое время и использовать в качестве исходных данных временные ряды.

3. Методики поиска ассоциативных правил, анализа исключений и секвенциального анализа

Для создания методики поиска ассоциативных правил, анализа исключений и секвенциального анализа необходимо рассмотреть существующие меры близости (эвклидово расстояние, расстояние по Хеммингу, расстояние Чебышева, расстояние Махаланобиса, пиковое расстояние), которые используются в перечисленных методах.

Евклидово расстояние в общем понимании широко используется в задачах кластерного анализа [6]. Мера близости определяется по формуле:

$$\rho_{\mathrm{E}}(x_i, x_j) = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2} \qquad (15)$$

где x_{ik}, x_{jk} — значения k-й характеристики у i-го (j-го) объекта.

Евклидово расстояние может быть применено в тех случаях, когда исходные характеристики не зависят друг от друга. Также данная мера близости не может применяться, когда исходные характеристики объектов неоднородны или по-разному важны для классификации.

В тех случаях, когда характеристики имеют разные единицы измерения, осуществляется нормирование каждой характеристики, что в свою очередь также может негативно повлиять на результат – если кластеры хорошо разделимы по одной характеристике и плохо разделимы по другой, то нормирование приведет к уменьшению классификационных возможностей первой характеристики.

Для случая, когда характеристики имеют разную степень важности, применяется «взвешенное» евклидово расстояние. Таким образом, каждой характеристике в соответствии с её степенью важности при классификации,

присваивается некоторый вес. Однако, определение степени важности подразумевает проведение дополнительных исследований.

Расстоянием по Хэммингу [7] называют меру близости, основанную на разности по координатам. В большинстве случаев данная мера близости приводит к таким же результатам, как и при использовании евклидова расстояния, однако для нее влияние отдельных больших разностей (выбросов) уменьшается. Это достигается тем, что данные выбросы не возводятся в квадрат. Расстояние по Хеммингу вычисляется по формуле:

$$\rho_H(x_i, x_j) = \sum_{k=1}^{N} |x_{ik} - x_{jk}| \qquad (16)$$

Расстояние Хэмминга применяется для строк одинаковой длины и служит функцией, определяющей расстояние в пространстве метрик объектов одинаковой размерности.

Расстоянием Чебышева [8] (или супремум-норма или метрика доминирования) между числовыми векторами называется максимум модуля разности компонент этих векторов. Данная мера близости используется в тех случаях, когда требуется определить различие объектами, причем для этого достаточно различия по какой-либо одной координате. Расстояние Чебышева определяется по формуле:

$$\rho_\infty(x_i, x_j) = \max_{1<k<n} |x_{ik} - x_{jk}| \qquad (17)$$

Расстояние Чебышева иногда используется в нейро-нечетких сетях. Основным недостатком данной меры близости является то, что полученные с помощью расстояния Чебышева кластеры пересекаются.

Расстояние Махаланобиса [9] является обобщением евклидова расстояния. Главное отличие от евклидова расстояния – учет корреляции

между характеристиками, а также независимость от масштаба. Расстояние Махаланобиса можно интерпретировать как степень расхождения между двумя случайными векторами из одного распределения вероятностей с матрицей ковариации S. Расстояние Махаланобиса вычисляется по формуле:

$$\rho_M(x_{i,}x_j) = (x_{i,} - x_j)S^{-1}(x_{i,} - x_j)^k \quad (18)$$

Данная мера близости плохо работает, если ковариационная матрица высчитывается на всем множестве входных данных. Тем не менее, в силу сосредоточенности на конкретной группе данных, данная мера близости показывает хорошие результаты

Пиковое расстояние [10] предполагает независимость между случайными переменными, что говорит о расстоянии в ортогональном пространстве. Но в практических приложениях эти переменные не являются независимыми. Пиковое расстояние определяется по формуле:

$$\rho_I(x_{i,}x_j) = -\frac{1}{m}\sum_{k=1}^{n}\frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}} \quad (19)$$

Для выбора меры близости необходимо обладать информацией о характере исходных данных. Так, обычное евклидово расстояние не подходит для кластеризации данных, измеренных в разных величинах. Нормализация с помощью деления на среднеквадратичное отклонение решает данную проблему, однако, это может также привести к усилению шумового эффекта некоторых характеристик. Также простое евклидово расстояние неприменимо, когда характеристикам необходимо задать определённые веса. Данный недостаток можно устранить, посредством использования «взвешенного» евклидова расстояния. Расстояние по Хэммингу позволяет снизить влияние выбросов на результаты

кластеризации, но данная мера близости неприменима, когда в исходных данных могут быть пропуски. Расстояние Чебышева хорошо подходит для кластеризации разряженных и зашумленных данных, когда для кластеризации достаточно установить различие хотя бы по одному из множества признаков. Однако для сложных данных такая мера близости неприменима. Расстояние Махаланобиса и пиковое расстояние не подходят для кластеризации больших объемов сложных данных, таких как временные ряды, для которых, как правило, и используются структуры типа OLAP.

Таким образом, рассмотренные меры близости могут быть использованы в различных системах кластерного анализа, в том числе и систем хранилищ с OLAP-структурой. Приведенные в разделе меры близости решают задачи кластеризации для различных наборов исходных данных, но не подходят для решения задач кластеризации больших объемов сложных данных, например, временных рядов. Для подобных задач больше подходят меры близости, основанные на коэффициентах корреляции или вейвлет-коэффициентах, рассмотренные в предыдущем разделе.

Для получения информации из данных OLAP-структур используются различные методы анализа. Обычно, эти методы заключаются в создании правил, на основе которых и выявляется информация. Найденные правила можно условно разделить на три вида:

- полезные правила – правила, которые легко обосновать, и на основе которых можно находить информацию с высокой энтропией.
- тривиальные правила - правила, которые легко обосновать, и на основе которых можно находить информацию с низкой энтропией.
- непонятные правила – правила, которые трудно обосновать, и на основе которых может быть получена как ложная информация, так и информация с высокой энтропией (скрытые знания).

Для повышения эффективности анализа OLAP хранилищ были разработаны три методики, которые позволяют выявлять правила для

исходного множества данных. Для данных, представленных стандартными множествами наборов, используется методика поиска ассоциативных правил. Для последовательностей наборов используется методика секвенциального анализа. Для поиска скрытых знаний и трактовки аномальных данных, используется методика анализа исключений. Рассмотрим данные методики более подробно.

Методика поиска ассоциативных правил

Количество элементов задается равным единице и выполняется подсчет поддержки для всех одноэлементных наборов. Данная операция проводится с целью выделения наборов, у которых значение поддержки превышает минимально заданное пользователем значение. Далее число элементов увеличивается на единицу. Если не удается создать наборы с требуемым количеством элементов, тогда алгоритм завершается. Если же удается, тогда из частых наборов создается множество наборов-кандидатов с текущим числом элементов (*n*). Для этого частые наборы с *n-1* числом элементов объединяются в кандидаты с числом элементов n. Кандидаты поочередно формируются с помощью добавления к одному частому набору с *n-1* числом элементов (*P*) элемента из другого частого набора с *n-1* числом элементов- *Q*. При этом добавляется последний элемент набора *Q*, имеющий высший порядок, чем последний элемент набора *P*, а все *n-2* элемента данных наборов являются одинаковыми.

Далее для каждой транзакции *T* из множества *D* выбирается кандидат C_t из множества потенциально частых наборов с текущим числом элементов (C_n), присутствующих в транзакции *T*. Для каждого набора из построенного множества C_n удаляется набор, если хотя бы одно из его *n-1* подмножеств не является часто встречающимся т.е. отсутствует во множестве L_{n-1}. Для каждого кандидата из C_n поддержка увеличивается на единицу. Затем кандидаты L_n выбираются из множества C_n, причем только те, у которых значение поддержки превышает заданную пользователем. По окончании данной операции необходимо снова увеличить количество элементов на

значение p, получаемое исходя из отношения (h_n) числа частых наборов к числу всех наборов-кандидатов. Так, если $h_n < 0.65$, то p риравнивается к единице. Если $0.65 < h_n < 0.7$, то $p=2$. В случае, если $h_n > 0.7$, количество элементов (значение длины) увеличивается на $p=3$. После проведения данной операции цикл повторяется, начиная с определения наборов-кандидатов с новым количеством элементов. Результатом работы алгоритма является объединение всех множеств L_n для всех n.

Методика секвенциального анализа

В рамках задачи секвенциального анализа требуется рассматривать уже не наборы элементов, а их последовательности. Для этой задачи предлагается использовать алгоритм AprioriDSP. Однако требуется учитывать дополнительные условия, чтобы определить, содержит ли последовательность указанную подпоследовательность. Условие формируется следующим образом: последовательность содержит подпоследовательность, если она содержит все элементы подпоследовательности. Также следует добавить к этому группировку данных. Так, транзакция T содержит объект X, если X присутствует в T, или X является потомком какого-либо элемента из T. Транзакция T содержит одноэлементную последовательность Y если в транзакции T присутствуют все объекты из Y. Также для анализа последовательностей вводится скользящее окно. При этом предполагается, что элемент последовательности может состоять из нескольких транзакций, если разница во времени между ними меньше чем размер окна. Бывает также необходимо ввести временные диапазоны для наборов, являющихся последовательностью упорядоченных элементов. Таким образом, условием, что последовательность $d = <d_1 \ldots d_k>$ содержит в себе подпоследовательность $s = <s_1 \ldots s_k>$, при заданных размерах окна, а также заданном диапазоне временного интервала между транзакциями, является существование таких целых чисел $l_1 \leq u_1 < l_2 \leq u_2 < \ldots < l_m \leq u_m$, что:

1) s_i содержится в $\bigcup_{k=l_i}^{u_i} d_k$, *при* $1 \leq i \leq m$;

2) разница во времени транзакций d_{u_i} и d_{l_i} меньше, либо совпадает с размером окна, при $1 \leq i \leq m$;
3) время транзакции d_{l_i} минус время транзакции $d_{u_{i-1}}$ больше минимального значения временного интервала, при $2 \leq i \leq m$;
4) время транзакции d_{u_i} минус время транзакции $d_{l_{i-1}} \leq$ максимального значения временного интервала, при $2 \leq i \leq m$.

Таким образом, методика секвенциального анализа состоит из следующих этапов:

1) определение содержания указанной подпоследовательности в последовательности упорядоченных элементов;
2) определение принадлежности элемента к транзакции с учетом таксономии;
3) формирование скользящего окна;
4) ввод временных диапазонов для наборов транзакций;
5) генерация кандидатов;
6) подсчет поддержки кандидатов.

Методика анализа исключений

Задача поиска исключений имеет большую важность с точки зрения интеллектуального анализа данных. Для снижения негативного влияния исключений на результаты анализа, задача сводится к определению, являются ли исключения результатами ошибок в исходных наборах данных, или же они являются отличающимися.Таким образом методику анализа исключений можно разделить на следующие этапы:

1) Определение множества наборов, являющихся исключениями.

На данном этапе к множеству наборов применяются правила, найденные алгоритмами поиска ассоциативных правил. Наборы, для которых данные правила не выполнялись, определяются как исключения.

2) Поиск новых правил для сформированного множества.

Определенные на предыдущем этапе исключения могут нести в себе скрытые правила. Для определения, являются ли данные наборы отличающимися, или же их можно отнести к ошибочным, для сформированного множества повторно применяются алгоритмы поиска ассоциативных правил. Для определения новых правил путем анализа наборов-исключений, следует задать значение минимальной поддержки и оценку достоверности исходя из специфических свойств исключений. Найденные правила требуют подтверждения, таким образом, правильность и эффективность процесса должны подвергаться экспертной оценке.

3) Очистка множеств, содержащих ошибки.

На данном этапе наборы, которые были определены как ошибочные (т.е. не содержащие скрытых правил) исключаются из множества.

4) Замена множеств в источниках.

Данный этап важен с точки зрения оптимизации дальнейшей работы с хранилищами данных. После того, как наборы-исключения, для которых не были найдены скрытые правила, были исключены, требуется заменить исходные множества, содержащие ошибки, очищенными. Таким образом, при дальнейшем извлечении данные не требуют повторной очистки.

Заключение

В рамках работы над данным проектом, все задачи запланированные на 2016 год были выполнены полностью, а именно:

1. исследование физического хранения на внешнем накопителе памяти реляционных, многомерных и гибридных хранилищ данных структуры «созвездие»:
    а) исследование физического хранения данных;
    б) исследование факта “взрыва” хранилища данных;

в) исследование оптимальной структуры хранения данных для выгрузки в системы интеллектуального анализа данных.

2. исследование мер близости временных рядов, основанных на коэффициенте корреляции и вейвлет-коэффициентах, их сравнительный анализ, а также, создание кластеризационно-корреляционного сканера хранилища данных структуры «созвездие»:
    а) разработка меры близости временных рядов, основанной на коэффициенте корреляции;
    б) исследование меры близости временных рядов, основанной на вейвлет-коэффициентах
    в) сравнительный анализ мер а) и б);
    г) разработка кластеризационно-корреляционный сканера хранилища данных.
3. Исследование возможности применения алгоритмов классификации, поиска ассоциативных правил, анализа исключений и секвенциального анализа к хранилищам данных структуры «созвездие».
    а) исследование прочих мер близости (эвклидово расстояние, расстояние по Хеммингу, расстояние Чебышева, расстояние Махаланобиса, пиковое расстояние);
    б) сравнительный анализ мер из предыдущего пункта с мерами близости основанными на коэффициенте корреляции и вейвлет коэффициентах;
    в) исследование возможности применения существующих алгоритмов классификации, поиска ассоциативных правил, анализа исключений и секвенциального анализа к хранилищам данных со структурой типа «созвездие».

Полученные результаты приведены в соответствующих разделах отчета:

1. Методика выбора сочетания реляционного и многомерного хранения данных (раздел 1);
2. Методика кластеризации временных рядов, полученных из хранилища данных, по мерам близости, основанным на коэффициенте корреляции и вейвлет-коэффициентах (раздел 2);
3. Методики поиска ассоциативных правил, анализа исключений и секвенциального анализа (раздел 3).

Кроме того, текущие результаты были представлены в журнале “Технические науки - от теории к практике” и на юбилейной XV Санкт-Петербургской Международной Конференции “Региональная информатика-2016” (“РИ-2016”). Библиографические ссылки:

- Михайлов М.В., Коломеец М.В., Булгаков М.В., Чечулин А.А. Исследование и определение основных достоинств и недостатков существующих типов хранилищ данных // Технические науки - от теории к практике: сб. ст. по матер. LXIV междунар. науч.-практ. конф. – Новосибирск: СибАК, № 59, 2016. С.22-27.
- Михайлов М.В., Чечулин А.А. Кластеризационно-корреляционный сканер хранилища данных // Юбилейная XV Санкт-Петербургская Международная Конференция “Региональная информатика-2016” (“РИ-2016”). 26-28 октября 2016 г. Материалы конференции. СПб., 2016. С. 181.

В рамках дальнейших исследований, планируется создание кластеризационно-корреляционного сканера, работающего за линейной время, зависящее от количества анализируемых объектов. Также планируется разработка методики адаптивного выбора методов классификации в зависимости от анализируемой OLAP- структуры.

СПИСОК ЛИТЕРАТУРЫ

1. *Михайлов М.В.* Инструментально-математические средства мониторинга страховой системы в режиме реального времени. Известия высших учебных заведений. Поволжский регион., No3(11), 2009, стр. 59 – 65.
2. *Михайлов М.В., Коломеец М.В., Булгаков М.В., Чечулин А.А.* Исследование и определение основных достоинств и недостатков существующих типов хранилищ данных // Технические науки - от теории к практике: сб. ст. по матер. LXIV междунар. науч.-практ. конф. – Новосибирск: СибАК, № 59, 2016. С.22-27.
3. *Компания SAP* Эффективные технологии построения корпоративных Хранилищ Данных – [Электронный ресурс] – Режим доступа – URL:http://www.olap.ru/desc/sybase/news/sybase.asp (Дата обращения: 20.11.2016).
4. *Бурнаев Е.В., Оленев Н.Н.* Меры близости на основе вейвлет коэффициентов для сравнения статистических и расчетных временных рядов // Межвуз. сб. научн. и научно-метод. трудов за 2005 г. г. Киров: Изд-во ВятГУ, 2006. Вып. 10. С.41-51.
5. *Нейский И.М.* Адаптивная кластеризация на основе дивизимных и итерационных методов // Информационные технологии в образовании, науке и производстве: сборник трудов третьей международной научно-практической конференции /под ред. Ю.А Романенко. М., 2009.
6. *Баргесян А.А., Куприянов М.С., Степаненко В.В.* Технологии анализа данных: Data Mining, Text Mining, Visual Mining, OLAP. Холод И.И. СПб.: изд. БХВ-Петербург, 2007. С.375
7. *Крянев А.В., Лукин Г.В.* Метрический анализ и обработка данных. Москва: изд. Физико-математическая наука, 2010. С. 280
8. *Крючников М.В.* Технология кластерного анализа финансовых показателей банков // Прикладная информатика. Вып. 1. Москва: изд. Синергия, 2006. С.41

9. *Баргесян А.А.* Анализ данных и процессов. СПб.: изд. БХВ-Петербург, 2009. С.512
10. *Манылов И.В.* Оценка точности распознавания классов при автоматизированной обработке аэрофотоснимков. // Известия высших учебных заведений. Приборостроение. Вып. 5 (54). СПб., 2011. С.35