УДК 004.6; 550.385.1

ГРУППИРОВКА БОЛЬШИХ ОБЪЕМОВ ИНФОРМАЦИИ ПО ВЫЧИСЛЯЕМЫМ ПРИЗНАКАМ С ПОМОЩЬЮ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

Ю.В. Пенских

Институт солнечно-земной физики СО РАН, Иркутск, Россия
penskikh@iszf.irk.ru

GROUPING A WEALTH OF INFORMATION BY CALCULATED SIGNS WITH THE USE OF RELATIONAL DATABASES

Yu.V. Penskikh

Institute of Solar-Terrestrial Physics SB RAS, Irkutsk, Russia

Аннотация. При работе с большими объемами информации часто возникает необходимость классифицировать данные на группы по каким-либо признакам, определить количество элементов в группе, их сумму, среднее значение и т.п. В данном исследовании задача классификации решалась с помощью реляционной базы данных. При этом группировка в базе данных не по скалярным, а по вычисляемым полям затруднительна. Для обхода этого ограничения был предложен способ предварительного расчета классифицирующих признаков, описаны преимущества использования данного подхода. Данная методика апробирована для группировки 250 суток минутных данных наземной сети магнитометров по узлам сетки в системе координат широта – местное геомагнитное время с одновременной группировкой по уровням активности AE-индекса.

Ключевые слова: группировка данных, база данных, вычисляемые признаки.

Abstract. When operating bulk information volumes, one often has to classify data into groups by some signs, to determine the number of elements in the group, their sum, mean value, etc. In this study, I solve the classification problem through a relational database. In doing so, it is difficult to group not by scalar, but by calculated fields within a database. To bypass this restriction, I propose a method to preliminary calculate the classifying signs, and describe the advantages of using this approach. This technique is tested for a group involving 250 days of 1-minute data from the ground-based network of magnetometers by hubs in the latitude-local geomagnetic time coordinate system with a simultaneous grouping by the AE-index activity levels.

Keywords: grouping data, database, calculated signs.

Введение

Практически в любой задаче требуется обработка и хранение данных. Файлы являются традиционным способом хранения информации на внешнем носителе. При всех своих достоинствах файлы имеют и ряд недостатков. Данные в файлах не контролируются операционной системой на логическую целостность, согласованность и непротиворечивость. Одновременное чтение-запись в файл затруднительно. Тяжело организовать работу с многотабличными сложноструктурированными взаимосвязанными данными. Не обеспечиваются даже простые операции, такие как суммирование и сортировка данных. Все эти задачи приходится реализовывать программно. Проблема усугубляется, если необходимо работать с файлами по сети и вести чтение-запись данных одновременно разными программами.

Данная проблема в первую очередь коснулась банковского сектора, для ее решения была создана реляционная модель данных [Codd, 1970]. В 1960–1980 гг. была создана теория реляционных баз данных, описывающая оптимальные принципы организации, структурирования и обработки информации, созданы реляционная алгебра и языки запросов к данным [Date, 2004]. Вскоре теорию воплотили в практику в виде реляционных систем управления базами данных (РСУБД). В настоящее время хранение информации в реляционных базах данных (БД) де-факто является стандартом для больших корпоративных приложений [Fowler, 2011].

В технике инверсии магнитограмм (ТИМ) долгое время основным хранилищем информации были именно файлы со всеми вышеописанными недостатками. За последние два года мы перешли на хранение геомагнитных данных мировой сети магнитометров, параметров солнечного ветра и др. в БД.

Цель настоящей работы состоит в упрощении задачи группировки больших объемов данных по вычисляемым признакам с помощью РСУБД и в приложении данного метода к геомагнитной БД.

Описание прикладной задачи

Для измерения магнитного поля Земли существует сеть наземных магнитометров. Постоянное увеличение объемов данных обеспечивает рост статистических исследований. На 2008–2009 г. работало около 288 станций [Gjerloev, 2012]. Их количество меняется в связи с вводом-выводом в эксплуатацию новых магнитометров, а так же с аварийной потерей данных наблюдений на существующих станциях. Сеть станций имеет неоднородное распределение на поверхности земли, что представляет трудность для задачи сферического гармонического анализа (СГА) [Базаржапов и др., 1979].

В методе ТИМ используется система координат дипольная широта (Ф) — местное геомагнитное

время (MLT). Эта система координат зависит от положения дипольного полюса, которое ежегодно меняется [Thébault et al, 2015]. В течение одной эпохи широта станции не меняется, а меняется только MLT. Так как каждая станция за сутки совершает один полный оборот, то после нормировки данных по MLT можно проводить статистические исследования (рис. 1).

Чтобы избежать увеличения количества уравнений в СГА и сделать данные относительно однородными, измерения группируются не по точным значениям Ф и MLT, а по однородной сетке. Исторически сложилось, что в ТИМ используется следующая сетка: шаг 1 градус по кошироте и 10 градусов по долготе (MLT). Каждое измерение относится к ближайшему узлу.

Для группировки по *AE*-индексу, измерения разбиваются на четыре класса: $AE \le 100$; $100 < AE \le 300$; $300 < AE \le 600$; $AE > 600$ [Spiro, 1982]. Положение станции в с.к. (Ф, MLT) зависит от мирового времени UT, географических координат станций и координат дипольных полюсов. При этом нас интересует одновременная группировка по нескольким признакам, т. е. измерения разбиваются на группы по *AE*, а затем на подгруппы по узлам (рис. 2).

В нашей задаче необходимо разбить на группы 250 сут минутных данных, выбранных из 2008 и 2009 гг. В среднем каждую минуту работает 215 магнитометров, которые выдают по три компоненты вариаций геомагнитного поля (рис. 3)

Рис. 1. Схема распределения измерений по узлам сетки

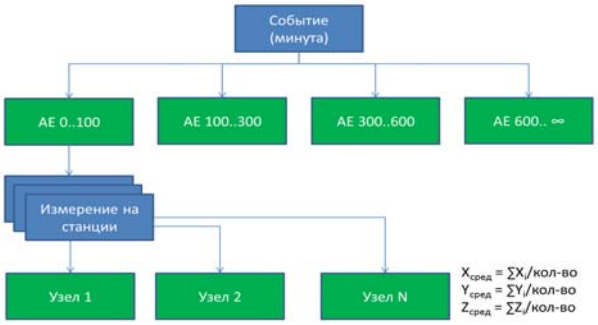

Рис. 2. Схема группировки данных

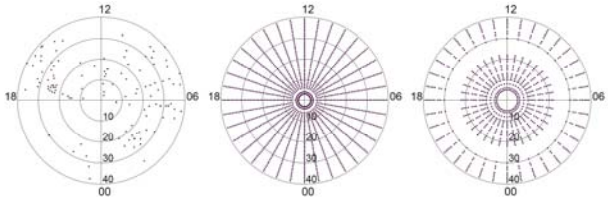

Рис. 3. Слева — распределение измерений станций северного полушария для 2008-03-08 14:00 UT; в середине — распределение измерений станций северного полушария за 250 сут при $AE \ge 600$; справа — распределение измерений станций южного полушария за 250 сут при $AE \ge 600$ (взгляд “сквозь землю”)

Программная реализация

Предварительно в БД загружаются географические координаты станций, координаты дипольных полюсов по эпохам, геомагнитные индексы и рассчитываются дипольные координаты станций в зависимости от эпохи. Перечисленные операции далее не участвуют в расчетах временных затрат.

Из исходного файла загружается суточный объем одноминутных вариаций геомагнитного поля. Для каждой минуты в момент загрузки выполняются следующие действия:

- преобразование в систему координат (Ф, MLT) с учетом эпохи;
- по времени определяется уровень *AE*, измерения *AE* классифицируется по группам *AE*;
- определяется узел (подгруппа), к которому относится измерение на станции; сохраняется сумма и количество попаданий в узел для каждой группы *AE*, узла и компоненты поля;
- по каждой группе, подгруппе и компоненте поля сумма делится на количество.

Данная реализация на Java обрабатывала суточный объем измерений за 10 мин, что, в целом, не плохо. Но на 250 сут потребовалось уже 1.74 сут. Программа получилась довольно узконаправленная, а добавление новых классифицирующих признаков потребовало бы внесения новых существенных изменений и еще больших временных и вычислительных затрат.

По полученным осредненным данным был проведен сферический гармонический анализ и построены эквивалентные токовые функции для каждого уровня *AE*-индекса. На рис. 4 представлена одна из таких функций.

Реализация с помощью БД

Для работы с БД используют SQL. SQL — язык программирования, применяемый для создания, модификации и управления данными. Одной из функций языка является группировка данных.

Примеры SQL запросов для схемы БД (рис. 5):

SELECT ae FROM kioto_ae — выбрать *AE* индекс из таблицы индексов геомагнитной активности.

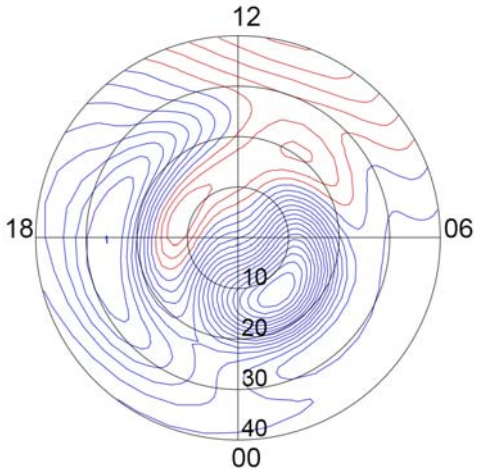

Рис. 4. Эквивалентная токовая функция, построенная по осредненным значениям для $AE \le 100$

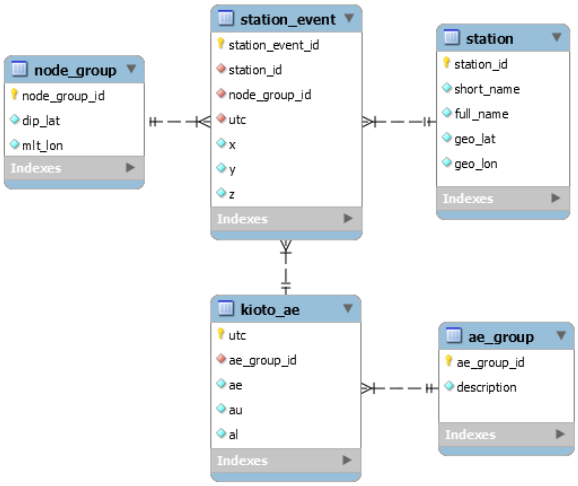

Рис. 5. Схема БД (демонстрационный вариант)

SELECT ae, COUNT(*), AVG(au), AVG(al) FROM kioto_ae GROUP BY ae — сгруппировать данные из таблицы по *AE* и вывести: *AE*, количество элементов в группе, среднее значение *AU* в группе, среднее значение *AL* в группе.

SELECT ae, au, COUNT(*), AVG(al) FROM kioto_ae GROUP BY ae, au — сгруппировать данные из таблицы по *AE*, *AU* и вывести: *AE*, *AU*, количество элементов в подгруппе, среднее значение *AL* в подгруппе.

Из примеров видно, что SQL позволяет легко группировать по одному, двум и более полям. Конструкция GROUP BY хорошо работает для группировки по полю, однако группировка по сложному вычисляемому признаку (функции) затруднительна. В SQL существует также конструкция HAVING, однако она имеет очень ограниченное применение. Для обхода данного ограничения SQL классифицирующие признаки были выделены в отдельные вспомогательные таблицы. БД денормализуется, так как появляются функциональные зависимости. Что, с одной стороны, противоречит принципу нормализации данных, с другой стороны, позволяет легко разрешить данный класс задач.

В данном случае для группировки по *AE* была создана таблица с группами по *AE* с полями: номер группы *AE*, описание группы (диапазон *AE*). Для группировки по MLT создана таблица: номер узла на сетке, коширота, долгота. Поле с номером группы *AE* было добавлено в таблицу с данными *AE*-индекса, а поле с номером узла на сетке было добавлено в таблицу с измерениями на станциях (рис. 5). С помощью такого метода можно сохранить информацию по сложным группирующим функциям в виде скаляров и проводить обычную группировку данных с помощью конструкции GROUP BY. Описанная методика сводит эту задачу к задаче группировки данных. БД выполняют группировку данных очень эффективно, т.к. они разрабатываются десятки лет сотнями высококвалифицированных специалистов.

Группировка измерений минутных вариаций геомагнитного поля за 250 сут была реализована с помощью БД Percona Server (www.percona.com). При этом программа на Java по загрузке суточного объема данных, пересчету компонент, расчету классифицирующих признаков и др. (как описано ранее), но без группировки данных, тратит всего 3 ч. Для группировки по рассчитанным признакам создается SQL запрос, который РСУБД выполняет всего за 2 мин.

Заключение

Задача группировки больших объемов информации по вычисляемым признакам существенно упрощается, если предварительно рассчитать классифицирующие признаки и использовать штатные средства СУБД.

Выполнен сравнительный анализ двух способов (программного и с помощью БД) решения задачи сортировки данных по двум признакам на примере 250 сут минутных вариаций геомагнитного поля.

Показана эффективность реализации с помощью БД.

Автор выражает благодарность С.Б. Лунюшкину за поставленную задачу и плодотворные дискуссии.

Работа выполнена при поддержке гранта РФФИ №15-05-05561.

Список литературы

Базаржапов А.Д., Матвеев М.И., Мишин В.М. Геомагнитные вариации и бури. Новосибирск: Наука, 1979. 26 с.

Code E.F. A Relational Model of Data for Large Shared Data Banks // Communications of the ACM. 1970. V. 13, no. 6. P. 377–378. DOI:10.1145/362384.362685.

Date C.J. An Introduction to Database Systems. – 8th edition, Addison-Wesley, 2003. 1024 p.

Fowler M. Patterns of Enterprise Application Architecture, Addison-Wesley. 2002. P. 33

Gjerloev J.W. The SuperMAG data processing technique. // J. Geophys. Res. 2012. V. 117, N A9. P. A09213. DOI:10.1029/2012JA017683.

Thébault E. et al. International Geomagnetic Reference Field: the 12th generation // Earth, Planets and Space. 2015. DOI: 10.1186/s40623-015-0228-9

Spiro R.W., Reiff P.H., Maher L.J. Precipitating electron energy flux and auroral zone conductances — An empirical model // J. Geophys. Res. 1982. V. 87, N A10. P. 8215–8227.