Методы оперативного анализа медико-демографических данных*

О. С. Исаева
Институт вычислительного моделирования СО РАН, Красноярск, Россия
e-mail: isaeva@icm.krasn.ru

В работе представлено решение задачи оперативного анализа медико-демографических данных. Механизм анализа данных основан на построении набора аналитических и статистических показателей. Описаны технологические особенности и проблемы, возникающие при построении хранилища данных и разработке алгоритмов анализа данных в многомерном представлении.

Ключевые слова: построение хранилищ данных, оперативная аналитическая обработка данных, унификация данных, агрегирование данных.

Введение

Технологии поддержки принятия управленческих решений в сфере здравоохранения должны основываться на достоверных и наглядных информационных моделях медико-демографических процессов. Выявление главных факторов смертности, анализ причинно-следственных связей величины и структуры смертности позволит адекватно распределять медицинские и административные усилия, обеспечит улучшение качества медицинской помощи и демографической ситуации в нашей стране.

В связи с возрастающими требованиями к оперативности и достоверности анализа информации, накапливаемой в разнородных источниках данных системы здравоохранения, актуальным направлением исследований во всем мире становится построение хранилищ данных (Data Warehouse) и развитие многомерных методов анализа, основанных на применении технологии оперативной аналитической обработки данных OLAP (On-Line Analytical Processing) [1]. Для построения хранилища медико-демографических данных и системы многомерного анализа выполнялось комплексное исследование предметной области, информационных источников и вычислительных алгоритмов.

Как правило, алгоритмы анализа медико-демографических данных работают с плоским представлением информации. При увеличении размерности задачи, усложнении представления данных стандартные методы анализа становятся непригодными и требуется адаптировать вычислительные алгоритмы к многомерной постановке задачи и работе с информационными гиперкубами. Основной задачей становится построение адекватной информационной модели многомерного анализа данных.

*Работа выполнена при финансовой поддержке гранта президента Российской Федерации для ведущих научных школ РФ № НШ-3431.2008.9.

В работе представлены технологические аспекты построения информационных моделей для анализа медико-демографических данных. Исследованы и описаны особенности и проблемы, возникающие при многомерной обработке данных.

Практическая реализация и выполнение исследований проводились с применением инструментальной системы "Менеджер хранилища данных" и профессиональной инструментальной среды объектно-ориентированной разработки OLAP-приложений системы "Аналитик" [2, 3] (разработки Института вычислительного моделирования СО РАН).

1. Задача анализа медико-демографических данных

Основу для изучения проблемы улучшения состояния здоровья населения в мировой практике составляют данные о случаях и причинах смерти. В нашей стране в условиях низкой рождаемости проблема снижения показателей смертности приобретает особую актуальность.

Механизм исследования медико-демографических данных основан на построении набора аналитических и статистических показателей. Основными показателями для измерения уровня здоровья населения являются: средняя ожидаемая продолжительность жизни при рождении; коэффициенты общей, повозрастной смертности, показатели смертности от диагностируемых заболеваний и др. [4]. Определение управляемых факторов смертности позволит адекватно распределять медицинские и административные усилия, обеспечить улучшение качества медицинской помощи и демографической ситуации.

Основная идея OLAP-технологии — представление информации в виде многомерных кубов, где оси определяют измерения, а в ячейках помещаются показатели. Выполняя операции агрегирования над гиперкубами данных, можно получить информацию требуемого уровня подробности. Специализированное хранилище данных — база, на которой строится аналитическая система, поэтому от способов организации данных и качества предобработки зависят функциональные возможности системы и удобство манипулирования информацией.

Цель работы: разработка и реализация информационных моделей для OLAP-анализа медико-демографических данных, которые позволят проводить статистический и оперативный анализ данных случаев смерти в динамике, вычислять агрегированные показатели, изучать структуру и тенденции показателей, выделять и анализировать факторы, влияющие на смертность населения, получать информацию в наглядной и доступной форме для определения комплекса мер, направленных на снижение смертности.

Под информационной моделью многомерного анализа в работе понимается формализованное описание информационных структур и операций над ними в понятиях и терминах OLAP. Для создания информационных моделей и реализации системы анализа медико-демографических показателей потребовалось решить ряд задач:

1) провести анализ предметной области: определить состав исходных данных, перечень показателей и алгоритмов их вычисления;

2) выполнить проектирование хранилища данных, разработать методы и средства его наполнения, актуализации, унификации и проверки достоверности данных;

3) создать методы аналитической обработки многомерных данных и алгоритмы анализа, учитывающие степень агрегации данных;

4) реализовать интерфейс и функциональное наполнение информационной системы с помощью инструментальных средств.

2. Построение специализированного хранилища данных

Хранилище данных — это база данных, поступающих от многих рабочих систем; данные интегрируются, собираются и структурируются таким образом, чтобы их можно было использовать в анализе и в процессе принятия управленческих решений [5].

Проектирование хранилища данных для анализа медико-демографических процессов заключается в создании структур таблиц фактов, т. е. в сведéнии данных из разных источников к общей структуре и в установлении связей с централизованно поддерживаемыми справочниками. Таблицы фактов содержат числовые значения показателей, по которым собирается статистическая информация: "Среднегодовая численность населения" и "Данные по умершим Красноярского края" [4].

Исходная информация имеет разные форматы (Excel, Access, Dbase, MsSql) и структуры хранения данных за разные периоды наблюдения. Анализ информационных источников показал, что основной набор данных можно разделить на три блока по типам источников.

В первый блок входят накопленные в Красноярском краевом медицинском информационно-аналитическом центре базы данных о случаях смерти. Эти данные физически лежат в разных базах данных, с отличающимися структурами таблиц и разной степенью детализации данных. Второй блок данных составляет демографический справочник. Он избыточен по сравнению с первым блоком данных, излишне детализирован и требует дополнительной подготовки для совместного анализа. Остальные специализированные справочники входят в третий блок данных.

Анализ предметной области показал, что, кроме изменения структур данных, меняются территориальные и медицинские справочники. Эти особенности приводят к невозможности динамического анализа данных без проведения дополнительных преобразований. Для заполнения хранилища потребовалось создать процедуры загрузки, преобразования, унификации и проверки корректности данных. Процесс унификации данных включал создание специальных справочников, таблиц соответствия и процедур преобразования данных.

В хранилище данных наряду с исходными помещаются предварительно агрегированные данные. Предварительное агрегирование данных обусловлено тем, что данные поступают в хранилище из разных источников и имеют разные уровни детализации, требуется группировка данных (например, по возрастным группам или по периодам времени), необходимых для реализации алгоритмов анализа.

В работе использовалась частичная предагрегация данных, определяемая в зависимости от предполагаемой при проектировании частоты использования данных. Выполнение полной предварительной агрегации данных в работе не используется. Число агрегатов зависит от количества детальных членов в измерениях и при полном агрегировании требуется построить порядка 2^m таблиц агрегатов, где m — число измерений. С ростом m увеличение числа таблиц агрегатов ведет к увеличению размера хранилища данных и к трудностям аналитической обработки [6].

Для удобства поиска и использования данных в OLAP-приложениях в репозитарии хранилища была выполнена семантическая группировка данных по степени детализации и способам агрегирования. Совместное хранение исходных и обработанных данных

позволяет проводить исследования с любой степенью детализации и в любом разрезе, но это требует создания специальных процедур для загрузки новых данных.

Когда в хранилище поступают новые данные, таблицы агрегатов оказываются не синхронизированы с таблицами данных. Данные пополняются ежегодно, и для сокращения трудозатрат по обработке новой информации созданы пакеты загрузки (в технологии ETL — extract, transform, load), которые являются совокупностью присоединенных процедур, выполняющихся в заданном порядке [7]. После загрузки выполняется многоступенчатая автоматическая обработка для частичного или полного обновления данных. Процедуры загрузки и обработки разделены на отдельные функции для обеспечения возможности встраивания на определенных этапах дополнительных процедур фильтрации или преобразования данных.

3. Проверка корректности данных

Основное требование аналитика — достоверность используемой информации [8]. Для контроля данных созданы алгоритмы пересчета вычисляемых полей, преобразования по таблицам соответствия; проверки типизированных значений, для которых задана область определения. На этапе проектирования хранилища методы проверки корректности данных не автоматизированы, их поиск и построение носят эмпирический характер.

Проверка проводилась как на соответствие отдельных значений измерений, так и на смысловую корректность наборов данных. Для этого средствами OLAP-анализа были построены многомерные гиперкубы с визуализацией в виде картограмм и диаграмм.

Для проверки корректности агрегированных по территории данных были рассмотрены диаграммы, отражающие динамику изменений показателей за периоды времени. Например, из построенной диаграммы численности населения (рис. 1) видно, что данные за разные годы существенно различаются. Для более детального анализа требуется

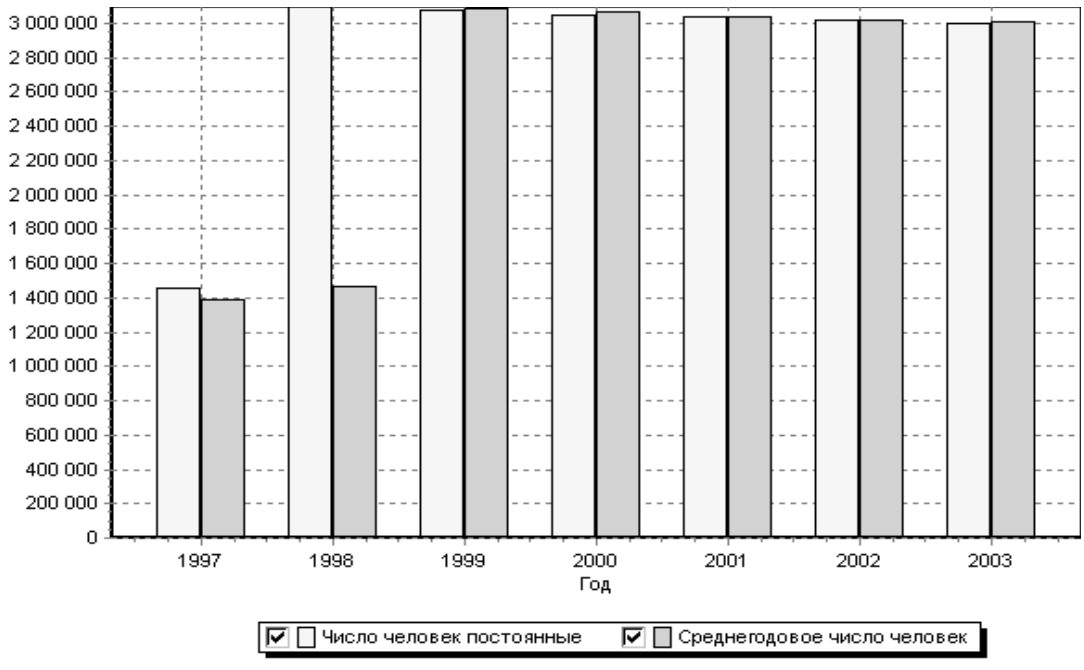

Рис. 1. Визуализация данных в виде диаграмм

Рис. 2. Картографическая визуализация данных

строить доверительные интервалы показателей. Такой подход применим в случае, когда понятно содержательное наполнение данных и можно представить достоверные значения показателей.

Еще один способ проверки корректности данных — картографическая визуализация. В картограммах территориальные объекты раскрашиваются в соответствии с цветами классов, построенных по значениям показателя. Так были построены картограммы (рис. 2) по показателям, относящимся к конкретному году наблюдений, что позволило обнаружить пропуски в данных.

Картографическая визуализация позволяет обращать внимание на показатели, принадлежащие соседним территориям, значения которых попадают в несмежные классы. Такие данные не обязательно свидетельствуют об ошибках и требуют отдельного исследования. Итерационный процесс загрузки, проверки корректности и обновления данных позволил повысить достоверность данных в хранилище.

4. Особенности построения аналитических моделей

Проектирование и реализация процессов взаимосвязи данных заключаются в построении показателей статистического и OLAP-анализа, учитывающих многомерное представление данных и результатов анализа. Информационные модели содержат аналитические показатели, построенные по международным и региональным методикам, адаптированным к технологии оперативного анализа данных.

В работе все алгоритмы расчета аналитических показателей построены так, чтобы аналитик в процессе исследования мог моделировать направление анализа: выбирать произвольные наборы данных, определять методы, рассматривать детальную ситуацию на конкретной территории или в масштабе региона. Показатели вычисляются непосредственно в момент обращения к ним, что обеспечивает автоматическое участие в

расчетах новых данных, импортируемых в хранилище, и позволяет избежать ошибок агрегирования.

При построении аналитических показателей возникают проблемы детализации и агрегирования данных, связанные с несбалансированностью данных и иерархий измерений. Измерения содержат внутренние смысловые иерархии, несбалансированность возникает при переходе от агрегированных данных к детальным, при этом появляются уровни, на которых отсутствует полный набор данных для расчета. Построенные алгоритмы анализа учитывают возможность отсутствия данных на разных уровнях иерархии и корректно обрабатывают такие ситуации или предлагают выполнять анализ в детальных точках, но за заданные периоды времени наблюдения.

Несбалансированность данных возникает из-за того, что таблицы фактов содержат разное число измерений. Если строить алгоритмы расчета, в которых используются показатели, заданные в одних и тех же разрезах анализа, то агрегирование данных при изменении набора измерений приводит к корректным результатам. В случае, если одни показатели наблюдались в большем числе разрезов, чем другие, при исключении этих измерений возникнут ошибки расчетов. Например, при построении аналитических показателей требуется в одном алгоритме использовать данные о численности населения и случаях смерти, которые помимо совместных измерений (“год”, “территория”, “возрастная группа”) имеют дополнительное измерение (“причины смерти”). Стандартно агрегирование данных при исключении измерений происходит суммированием всех значений, но в нашем случае это приведет к многократному увеличению (рис. 3) показателя численности населения по числу значений исключенного измерения. Для расчета таких показателей указывается способ агрегирования, исключающий из суммирования данные по не относящимся к нему измерениям.

При анализе с высокой степенью детализации данных появляются проблемы, связанные с разреженностью таблиц. Разреженность возникает при объединении гиперкубов с данными из разных таблиц по одному набору измерений. Для решения этой проблемы устанавливается тип соединения таблиц (внутреннее, внешнее и пр.) в зависимости от измерений данных, требуемых для реализации алгоритма анализа.

Выбор методов статистического анализа также связан с особенностями многомерности данных. В постановке OLAP важно найти баланс между упрощением вычислительных алгоритмов (для сокращения времени получения результатов) и обеспечением корректности расчетов при детализации данных [8]. В системе построены показатели смертности, а также стандартное отклонение и доверительные интервалы, характеризующие достоверность (надежность) изменения смертности.

Рассматривается следующая постановка задачи: выполнено n независимых испытаний, в которых некоторое событие произошло x раз. Требуется найти интервальную оценку неизвестной вероятности появления события, которое имеет два исхода “умер”, “не умер”. Для построения доверительных интервалов используется биномиальное рас-

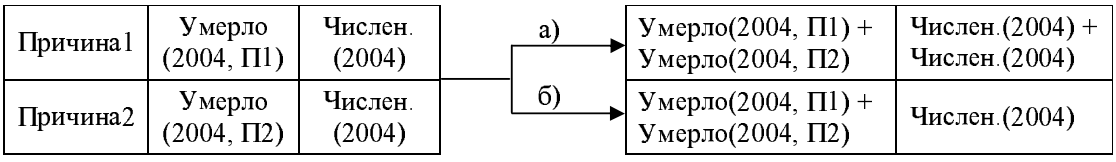

Рис. 3. Пример агрегирования данных

пределение случайной величины $B : n, p, f(x) = C_n^x p^x q^{n-x}$, где p — вероятность реализации первого исхода, $q = (1-p)$ — второго исхода, n — число испытаний (численность населения в заданной группе). При определенных значениях n и p это распределение аппроксимируется нормальным распределением, для которого существуют достаточно простые формулы вычисления доверительных интервалов. Применение таких формул позволяет строить быстрые вычислительные алгоритмы. В случае, если такой аппроксимации нет, то границы доверительного интервала рассчитываются более сложным образом:

$$\theta_{\text{лев}} = \sum_{i=0}^{x} C_n^i p^i q^{n-i} < \frac{\varepsilon}{2}, \quad \theta_{\text{прав}} = \sum_{j=n}^{x} C_n^j p^j q^{n-j} < \frac{\varepsilon}{2}.$$

Исследование показало, что для данных по крупным территориям (с большой численностью населения) критерии аппроксимации выполняются, для детальных данных — нет. Реализованные в работе алгоритмы учитывают степень детализации данных.

Построение сложных аналитических алгоритмов в многомерной постановке требует применения специальных технологических подходов. Рассмотрим пример построения "Таблиц дожития" по методике ВОЗ, описанной в [9]. Таблицы дожития содержат результаты расчета ожидаемой продолжительности жизни — число лет, которое в среднем предстоит прожить поколению родившихся в определенном году, при условии, что на протяжении жизни поколение сохраняет повозрастные показатели смертности данного года.

В расчете используется большое число промежуточных вычислений. Гиперкуб исходных данных формируется с несколькими измерениями ("год", "район", "пол", "тип поселения" и пр.) и двумя показателями — "численность населения" и "количество умерших". Анализ выполняется с данными, расположенными на грани гиперкуба, построенной по измерению "возрастные группы". Часть показателей вычисляется в прямом направлении по оси "возрастные группы" (от младших к старшим), другая часть — в обратном направлении, при этом используются результаты расчета, полученные на предыдущих шагах.

В одной витрине данных, вычисляя показатели последовательно, решить эту задачу не удается из-за смены направлений прохода по оси возрастных групп. Первое и стандартно применяемое решение заключается в делении вычислительного алгоритма на части, содержащие показатели, которые могут строиться совместно. В нашей задаче расчет потребовалось разделить на две подзадачи, решаемые отдельно. Такой подход оправдан в случае, если мы не имеем дело с многомерными данными.

При OLAP-анализе аналитик выделяет подмножество данных с одной фиксированной осью и остальными произвольными измерениями, т. е. получает грань гиперкуба. Применяется первый алгоритм, промежуточный результат сохраняется в виде таблицы агрегатов. К исходному набору данных (с фиксированным числом измерений) и таблице агрегатов применяется второй алгоритм анализа для получения окончательного результата. Выбирается следующий набор измерений и повторяется вся последовательность действий. Таким образом, строятся набор граней гиперкуба и набор показателей, которые корректны при заданном наборе измерений. Такой пошаговый подход ведет к увеличению вычислительной сложности и росту числа промежуточных таблиц-агрегатов. Например, если одно измерение формирует набор граней для алгоритма расчета и существует n независимых измерений, то число возможных наборов исходных данных равно 2^n и требуется написать при разработке моделей один программный алгоритм

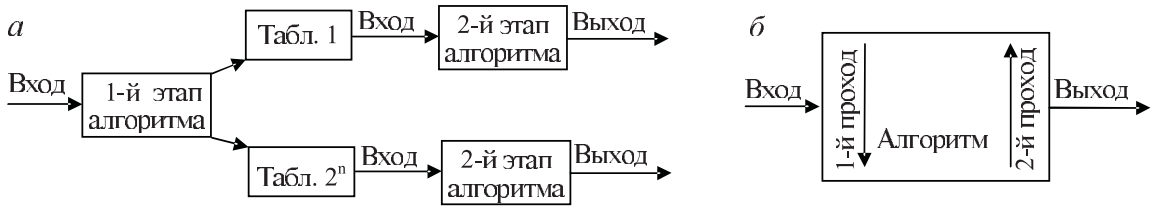

Рис. 4. Стандартный алгоритм расчета (*а*); многомерный алгоритм (*б*)

1-го этапа и 2^n алгоритмов 2-го этапа расчета (рис. 4). В случае группировки измерений, учитывающей внутреннюю иерархию, невозможно предсказать на этапе создания моделей, какие наборы данных для анализа построит аналитик.

Во избежание представленных недостатков расчет в работе выполняется так, чтобы набор данных всегда оставался значимым для алгоритма. Для этого применяются средства организации и контроля разнонаправленных проходов вычислительных процедур. Для данной задачи построен двухпроходный алгоритм расчета. В программной реализации введены переменные, отвечающие за идентификацию координат набора данных по измерениям. Логика работы, заложенная в алгоритм, отслеживает переход к новой грани куба. При таком подходе, реализовав небольшое число алгоритмов, можно проводить большой спектр исследований и свести процесс анализа к выбору набора измерений и вызову расчета показателей, получая результаты за определенный год или в динамике по годам, по отдельным районам региона или по всей территории в целом.

Заключение

В работе представлены этапы и технологические особенности оперативной аналитической обработки медико-демографических данных. Результатом работы стали алгоритмы и методы наполнения, унификации, актуализации специализированного хранилища данных, поступающих из разнородных оперативных баз, а также алгоритмы и методы статистического и OLAP-анализа, учитывающие многомерное представление данных.

Унификация и проверка корректности позволяют повысить достоверность и сопоставимость данных, что определяет высокую степень их информативности. Частичная агрегация данных на этапе проектирования хранилища, основанная на исследовании области применения данных, позволяет повысить оперативность получения результатов анализа.

Предложенные методы реализации алгоритмов анализа данных позволяют решать проблемы детализации и агрегирования данных, связанные с их несбалансированностью, иерархией измерений, и сокращать объемы вычислений при статистическом анализе данных и организации сложных многошаговых вычислений.

Поставленные в исследовании задачи решены, практическим результатом работы стала информационная система “Анализ медико-демографических процессов”. Она предназначена для проведения статистического и OLAP-анализа медико-демографических данных с получением результатов в наглядной и доступной форме для выявления факторов риска и определения комплекса мер, направленных на снижение показателей смертности.

Список литературы

[1] Хоббс Л., Хилсон С., Лоуенд Ш. Разработка и эксплуатация хранилищ баз данных. М.: КУДИЦ-ОБРАЗ, 2004. 586 с.

[2] Жучков Д.В., Кардашов Д.В. Программные средства поддержки централизованного хранилища медицинской информации // Тр. Всерос. конф. "Информационно-аналитические системы и технологии в здравоохранении и ОМС". Красноярск: КМИАЦ, 2002. С. 237–245.

[3] Горохова А.В., Ишенин П.П., Никитина М.И. OLAP-средства системы "Аналитик" // Труды Всерос. конф. "Информационно-аналитические системы и технологии в здравоохранении и ОМС". Красноярск: КМИАЦ, 2002. С. 220–228.

[4] Виноградов К.А., Денисов В.С. К проблеме исследования смертности населения в регионе с различной плотностью населения // Сб. научных трудов "Актуальные вопросы здравоохранения и медицинской науки". Красноярск, 2001. Вып. 2. С. 76–77.

[5] Стулов А. Особенности построения информационных хранилищ // Открытые системы. 2003. № 4.

[6] Хрусталев Е.М. Агрегация данных в OLAP-кубах // Алеф Консалтинг & Софт, 2003.

[7] Imhoff C. Understanding the Three E's of Integration EAI, EII and ETL Intelligent Solutions // DM Review Magazine, 2005.

[8] Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. СПб.: БХВ-Петербург, 2004. 336 с.

[9] Медков В.М. Демография: учеб. пособие. Ростов н/Д: Феникс, 2002. 448 с.

Поступила в редакцию 11 марта 2008 г.,
в переработанном виде — 3 июня 2008 г.