2. Алибеков А.К., Горшков В. В. К вопросу оптимизации параметров каналов. Обеспечение охраны, улучшения и восстановления поверхностных водных объектов в Западно-Каспийском бассейновом округе: Сб.статеймежрегион. науч.-практ.конференции. – Пятигорск: Западно-Каспийское бассейновое водное управление, 2011. – С. 179 – 182.
3. СНиП 2.06.03-85 Мелиоративные системы и сооружения. - М.: ЦИТП Госстроя СССР, 1986. – 65 с.
4. Учинчус А.А. Гидравлические и технико-экономические расчеты каналов. М.: Изд-во лит-ры по строит-ву, 1965. -274 с.
5. Чугаев Р. Р. Гидравлика. – Л.: Энергия, 1982. – 672 с.

УДК 528.854.2/ ББК 32.813

Бутенков С.А, Бесланеев З.О., Нагоров А.Л.

ГЕОМЕТРИЧЕСКИЙ ПОДХОД К ПОСТРОЕНИЮ МОДЕЛЕЙ ДАННЫХ НА ОСНОВЕ ТЕОРИИ ГРАНУЛЯЦИИ

Butenkov S.A., Beslaneev Z.O., Nagorov A.L.

GEOMETRICAL APPROACH TO THE DATA MODELS DESIGN, BASED ON THE GRANULATION THEORY

В настоящее время широко используются базы данных (БД), основанные на логических моделях реляционной алгебры. Однако в случае графических БД такие модели приводят к необходимости использования значительных ресурсов хранения. Развиваемый в работе геометрический подход, в сочетании с идеологией грануляции многомерных данных, предлагает методологию построения графических БД со сжатием данных за счет грануляции, примененную в работе к графическим БД изображений объектов сложной формы.

Ключевые слова: графическая база данных, информационная грануляция, информационная оптимизация.

A different kinds of databases, based on the relational algebra models, are used at present time. In case of graphical databases the known models are need of very large information resources for the data store and data management. In this paper a geometrical approach to the data granulation presented. As a practical result of granulation very sufficient data compression is obtained. The practical application of the new kind of data models is the image databases for the complicated form objects.

Key words: *graphical database, information granulation, information optimization.*

Введение

В работах Е.Ф. Codd и его последователей была создана теория моделирования баз дискретных данных, основанная на использовании математического аппарата реляционной алгебры (relational algebra) [1]. Эта математическая теория применяется для построения логических моделей, на основе которых в настоящее время построены широко распространенные системы управления реляционными БД (СУБД) [2]. Иной подход к проблеме передачи информации с помощью отношений предлагается в работах L. Zadeh [4-6]. В этих работах тоже рассматриваются методы моделирования данных на основе общей теории информационной грануляции (ТИГ), введенной L. Zadeh. Важнейшей особенностью ТИГ, которая явно не отражена в моделях реляционной алгебры, является возможность грануляции (укрупнения) данных в базе [5]. Идея грануляции основана на использовании свойства неразличимости некоторых наборов данных [7].

Использование методов геометрии (в общем случае – топологии) позволяет сжимать данные при хранении в БД путем грануляции, значительно уменьшая объем хранимых данных БД [8]. Методология, позволяющая применять грануляцию в графических БД, основана на применении общей ТИГ L. Zadeh к многомерным данным. Она развита в работах [7-9]. В графических БД нового типа хранятся не кортежи, а гранулы данных, число которых может быть значительно меньше, чем полное число реляционных кортежей графических данных [8].

Постановка задачи

Рассмотрим основные положения теории информационной грануляции, чтобы на их основе сформулировать задачу разработки методологии, позволяющей строить гранулированные графические БД.

В ТИГ *информационной гранулой* называется подмножество универсума U , на котором определено отношение сходства, неразличимости и т.п. [5]. В соответствии с основными положениями ТИГ, пусть ${}^1G_n, \dots, {}^kG_n$ – гранулы в универсумах ${}^1U_n, \dots, {}^kU_n$ для произвольной размерности n соответственно, тогда гранула, определяемая декартовым произведением $G_n = {}^1G_n \times \dots \times {}^kG_n$, называется *декартовой гранулой*. Множество гранул, которое содержит все объекты универсума, называется *гранулированием универсума*. Подмножество $A \subseteq U$ называется *составной гранулой*, если оно представляет собой дизъюнкцию (1) атомарных декартовых гранул. Согласно ТИГ, задав на осях (домены данных), обозначаемые в ТИГ как ${}^1P = \text{proj}_1 G_2$ и ${}^2P = \text{proj}_2 G_2$, мы тем самым определим декартову гранулу G_2 в виде $G_2 = {}^1P \times {}^2P$. (см. Рис. 1а) [6].

Рисунок1 - Декартова гранула на плоскости G_2 (а) и инкапсулирующая декартова гранула G_2^+ (b) по [4].

Для произвольной (не декартовой) гранулы определим декартову гранулу G_2^+ , называемую *инкапсулирующей гранулой* для G_2 , по следующему правилу:

$$G_2^+ = P \times^2 P, \mu_{1_P}(^1x) = \sup_{^2x} \mu_{G_2}(^1x, ^2x), \mu_{2_P}(^2x) = \sup_{^1x} \mu_{G_2}(^1x, ^2x), ^1x \in X, ^2x \in Y. \quad (1)$$

Рис. 1b иллюстрирует тот факт, что гранула G_2^+ является *точной верхней гранью* декартовых гранул, которые содержат G_2 [6]. Процедура покрытия исходных данных декартовыми гранулами по (4) называется в ТИГ *инкапсуляцией данных* [4]. Приближенное представление исходных данных в форме (1) с использованием инкапсулирующих декартовых гранул представляет собой *каноническую форму* представления многомерных данных [7] и реализует обобщенные ограничения $r = g$ или $r = fg$ (Рис. 1b).

С геометрической точки зрения модель (1) на плоскости строится на основе трех точек, образующих декартову гранулу G_2 (или $(n+1)$ точки для размерности n). Важнейшим качеством этой модели является то, что на основе (1) можно построить *модель инкапсулирующей гранулы* G_2^+ [8]. Например, для двух произвольных *непересекающихся* гранул на плоскости G_2^i и G_2^j модель инкапсулирующей гранулы G_2^+ представляется в весьма простой форме:

$$G_2^+(G_2^i, G_2^j) = \begin{pmatrix} \min(^1x_1^i, ^1x_1^j) & \min(^2x_1^i, ^2x_1^j) & 1 \\ \max(^1x_2^i, ^1x_2^j) & \max(^2x_2^i, ^2x_2^j) & 1 \\ \max(^1x_3^i, ^1x_3^j) & \min(^2x_3^i, ^2x_3^j) & 1 \end{pmatrix}. \quad (2)$$

Все приведенные выше определения для гранул на плоскости легко обобщаются на случай пространства данных произвольной размерности n [9].

Практической целью введения представление графических данных (изображений) в виде гранул является существенное уменьшению объема

хранимых гранулированных данных [7]. Разумеется, сжатие объема хранимых данных достигается за счет потери информации при гранулировании, но этот процесс можно оптимизировать, вводя числовые критерии качества представления данных согласно [8].

Для формулировки задач настоящей работы сформулируем основные вопросы применения базовых понятий ТИГ к *многомерным данным*. Во-первых, какова каноническая форма многомерных данных, т.е. форма, которая разъясняет смысл (для многомерных данных – геометрический смысл) обрабатываемых данных? Во-вторых, каким образом можно ввести метод приближенного представления многомерных данных, поскольку их прямое описание средствами аналитической геометрии может оказаться невозможным [7].

Рассматривая первый вопрос, можно показать, что, в согласии с идеями реляционной алгебры [2] и с идеями ТИГ по L. Zadeh [5], геометрические данные представляются в виде произвольных подмножеств декартовых произведений доменов на осях данных (атрибутов в терминологии реляционной алгебры). Это аналог исходных, неразъясненных данных (в терминологии ТИГ), т.е. данных, не приведенных к канонической форме. Приведение исходных данных к канонической форме разъясняет их смысл (в случае многомерных данных – геометрический смысл). Применительно к специфике графических БД, появляется возможность приведения разнородных графических данных к единой (канонической) форме, которую можно использовать для хранения в графической БД. В качестве дополнительного практически важного требования примем условие возможности компактного представления канонической формы данных, т.е. сжатия данных при хранении в БД. Отметим, что этим требованиям удовлетворяет гранулированное представление многомерных данных, введенное в работах [4,7,8]. Используем математические методы теории грануляции многомерных данных для построения гранулированных графических БД, для чего рассмотрим теоретическую базу оптимальной грануляции графических данных.

Оптимальная грануляция графических данных

С точки зрения проектирования моделей БД наиболее «прозрачны» энтропийные критерии гранулирования, связанные с информационным содержанием графических данных. Используем методологию построения критериев, предложенную в [7,8], используя тот факт, что модель (2) позволяет вычислять меры на моделируемых гранулах. В исходных графических данных (Рис. 1б) можно выделить, как минимум два подмножества, условно называемые *объект* и *фон* [7]. В ТИГ изображение есть множество двумерных подмножеств, одни из которых принадлежат объекту, а другие – фону. Подобное (бинарное) представление графических данных является крайне важным для задач анализа и

классификации данных [7]. Рассмотрим дискретные бинарные данные, структурированные на объект (*Object*) и фон (*Background*). Пусть данные содержат K информационных элементов и L объектов. Тогда взаимно обратные вероятности принадлежности случайно выбранного пиксела к одному из L объектов или к фону запишем в виде

$$p_O = L/K, \quad p_B = (K-L)/K = 1 - p_O. \quad (3)$$

Полная энтропия исходных дискретных данных, запишется как

$$H^n = H_O + H_B = -p_O \log_2(p_O) - p_B \log_2(p_B) = -p_1 \log_2((1-p_1)/p_1) - \log_2(1-p_1). \quad (4)$$

Оценить изменение полной энтропии (4) за счет покрытия бинарных данных декартовыми гранулами (2) (см. Рис. 1b) мы можем с помощью мер на гранулах покрытия. Пусть v – количество n -мерных гранул типа (2), принадлежащих универсуму данных U_n . Тогда для общего случая n -мерных бинарных данных мы можем вычислить полную энтропию гранулированного представления исходных данных H_G^n в виде:

$$H_G^n = \frac{\sum_{i=1}^v \eta_{G_n^i}^3}{|U_n|} \log_2 \left(\left(1 - \frac{\sum_{i=1}^v \eta_{G_n^i}^3}{|U_n|} \right) / \left(\frac{\sum_{i=1}^v \eta_{G_n^i}^3}{|U_n|} \right) \right) - \log_2 \left(1 - \frac{\sum_{i=1}^v \eta_{G_n^i}^3}{|U_n|} \right). \quad (5)$$

Таким образом, оптимизация информационного содержания покрытия многомерных графических данных декартовыми гранулами сводится к оптимизации информационного критерия, вычисляемого по покрытию данных n -мерными гранулами (5):

$$J = \min_{\bigcup_{i=1}^v G_n^i} (H^n - H_G^n). \quad (6)$$

Авторами предложены различные численные методы, реализующие вычисление (5) для оптимизации (6). Введенная модель может быть легко реализована в виде логической модели средствами построения реляционных БД, описанными в работах [2,3,8] и реализованными в комплексе программ управления графической БД [10]. Изучим свойства предложенной инфологической и логической модели [8] на примере графической БД изображений объектов сложной формы.

Результаты экспериментального изучения графической БД изображений

В качестве исходных данных для графической БД использовались изображения снежинок, накапливаемые в базе для дальнейшей их морфологической классификации [7]. В реляционных графических БД чаще всего используется векторное представление изображений на основе

стандартных примитивов [2,3]. Трудность применения векторного представления в проектируемой графической БД состоит в том, что форма снежинок неповторима. В результате становится невозможным использование стандартных векторных примитивов при представлении изображения. Рассматривая в качестве альтернативы растровый метод хранения данных, мы приходим к необходимости использования очень большого объема данных (для хранения всех пикселей изображения, включая неинформативные [7]). Проблема уменьшения объема хранимых данных в графической БД изображений снежинок решается путем использования метода грануляции многомерных данных [9].

Для реализации и исследования свойств гранулированной модели графической БД изображений объектов сложной формы был разработан программный комплекс, позволяющий экспериментально исследовать все основные характеристики графических БД [10]. Он оперирует бинарными изображениями объектов, образцы которых представлены на следующем рисунке.

Рис. 2 демонстрирует примеры покрытия исходных изображений объектов декартовыми гранулами в сравнении с Рис. 1. Отметим, что для наглядности на данном рисунке размеры гранул значительно увеличены в сравнении с оптимальными размерами по критерию (6).

Рисунок 2 - Пример гранулированного представления изображений в графической БД.

Основным параметром реализации графической модели является число разбиений изображения по осям (с помощью наносимой сетки по Рис. 3). Этот параметр определяет достигаемый коэффициент сжатия графических данных и число хранимых в графической БД гранул (см. следующий рисунок).

Рисунок 3 - Результаты экспериментального исследования параметров грануляции изображений сложной формы в графической БД

Изучение Рис. 3 показывает, что для заданного класса изображений коэффициент сжатия данных может достигать 70-80. При этом число хранимых гранул (связанное с ним приближенно обратной зависимостью) не превышает нескольких сотен. Это существенно меньше общего числа пикселей изображений, хранимых в графических БД с растровым представлением. Качество сжатия хранимых графических данных (потеря исходной информации (5)) оценивается числовым энтропийным критерием (6).

Заключение

Предложенный в работе грануляционный подход к построению графических БД, основанный на геометрии примитивных элементов (гранул), синтезирует идеи пространственной грануляции многомерных данных по [7] и реляционной алгебры [1]. Это позволяет вместо параметров отдельных точек растра хранить в графических БД параметры гранул, представляющих собой покрытия подмножеств элементов графических данных. В результате удастся в десятки и сотни раз снизить объем хранимых данных в графических БД. Отметим также, что при выборке графических данных из БД не требуется их декодирование из формата хранения, как в методах, использующих алгоритмы LZV и подобные ему. Новый подход позволяет также развивать полученные математические модели данных с целью использования криволинейных координат, наиболее подходящих для различных типов графических данных (цветные изображения и различные виды специальных изображений). Большие преимущества в организации структуры графических БД обещает также применение в гранулированных графических БД инфологических моделей, предложенных в [10].

Библиографический список:

1. Codd E.F. A relational model for data for large shared data banks / CACM, 1970.

2. Тиори Т., Фрай Дж. Проектирование структур баз данных. В 2 кн., – М.: Мир, 1985. Кн. 1. – 287 с.: Кн. 2. – 320 с.
3. Препарата Ф., Шеймос М. Вычислительная геометрия: Введение. Под ред. Ю. М. Банковского. – М.: Мир, 1989. – 478 с.
4. Zadeh L.A. Fuzzy sets and information granularity // in: Advances in Fuzzy Set Theory and Applications, Gupta, N., Ragade, R. and Yager, R. (Eds.), North- Holland, Amsterdam, 1979, pp. 3-18.
5. Zadeh L.A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic // Fuzzy Sets and Systems, vol. 90, p. 111–127, 1997.
6. Zadeh L.A. From Computing with Numbers to Computing with Words – From Manipulation of Measurements to Manipulation of Perceptions// IEEE Transactions on Circuits and Systems. – 1999. – vol.45. – P.105- 119.
7. Butenkov S. Granular Computing in Image Processing and Understanding // in: Proceedings of International Conference on Artificial Intelligence AIA-2004, Innsbruck, 2004, pp. 811-816.
8. Бутенков С.А., Кривша В.В., Бутенков Д.С. Гранулированные вычисления в системах интеллектуального анализа пространственных данных // В сб. трудов Международной конференции “Интеллектуальный анализ информации 2005” (ИАИ-2005), Киев, 2005, с. 108-117.
9. Бутенков С.А. Алгебраические модели в задачах интеллектуального анализа многомерных данных // Сб. трудов международной научно-технической конференции “Математическая теория систем 2009” (МТС-2009), Москва, 26-30 января 2009, с. 93-101.
10. Рогозов Ю.И., Бутенков С.А., Кодачигов В.И., Микита Р.М., Свиридов А.С. Информационные ER++ модели – новый подход к интеграции основных этапов проектирования информационных систем // Известия ТРТУ (технические науки), №9, 2006, с. 70-74.