«Оптика атмосферы и океана», **18**, № 9 (2005)

УДК 535.343.4+004.78

А.В. Козодоев, А.З. Фазлиев

Информационная система для решения задач молекулярной спектроскопии. 2. Операции преобразования наборов параметров спектральных линий

Институт оптики атмосферы СО РАН, г. Томск

Поступила в редакцию 18.07.2005 г.

Накапливаемые в банках данных HITRAN и GEISA параметры спектральных линий используют формат записи данных в виде строки. Такая информационная структура удобна для использования в приложениях, но практически неэффективна для использования в информационных системах и неприменима для машинной обработки данных. Предлагается дополнить структуру параметров спектральных линий тремя атрибутами для некоторых физических величин и значениями полуширин линий, обусловленных столкновениями с такими буферными газами, как неон, пары воды, углекислый газ и т.д. Для формирования составных банков данных предложен набор операций и описана их реализация на наборах параметров, хранящихся в базе данных.

Введение

В настоящее время развитие молекулярной спектроскопии сопровождается накоплением значительных объемов данных, в первую очередь параметров спектральных линий. Трудоемкость эксперимента и расчета приводит к тому, что отдельные лаборатории, проводя вычисления или измерения физических величин, занимаются изучением свойств лишь нескольких молекул. Как правило, даже при исследовании ограниченного числа молекул удается определять лишь часть набора параметров спектральных линий. Полученные данные публикуются в виде отдельных статей.

На практике все шире используются публикации в электронных версиях журналов, содержащие приложения с численными значениями параметров спектральных линий. Большая часть данных в них представляется в формате, принятом в банке данных HITRAN [1]. Такой подход существенно упростил процедуру тиражирования данных. Тот факт, что значительная часть программ, созданных ранее, использует этот формат, обусловлен традициями и некоторыми особенностями языков программирования, на которых ранее создавались вычислительные приложения.

Включение в информационную систему [2–4] типовых спектроскопических вычислительных программ в качестве сервисов заставляет по иному рассматривать требования к структурированию банка параметров спектральных линий. Существуют две причины, по которым необходимо менять структуру формата, предложенную создателями HITRAN'а:

во-первых, этот формат не является расширяемым (не предусматривает возможности введения новых физических величин, например полуширины линии, обусловленной столкновениями с инертными газами);

во-вторых, машинная обработка атрибутов физических величин, входящих в него, требует значительных дополнительных усилий.

Стоит отметить необходимость формирования метаданных при работе с параметрами спектральных линий. На наш взгляд, особую роль здесь начинают играть метаданные, описывающие операции, проводимые над данными. Такой тип метаданных-скриптов [5] только начинает развиваться.

Предлагаемый нами подход к структурированию параметров спектральных линий для машинного обмена их наборами и ввода этих наборов в информационные системы позволяет избежать непроизводительных затрат, обусловленных, например, работой с разными версиями банков данных HITRAN и GEISA. Стоит упомянуть, что объединение спектральных данных в таких банках данных, как HITRAN или GEISA, происходит не чаще чем раз в два или три года. Таким образом, на наш взгляд, существует потребность в создании средств для сбора и компоновки составных банков параметров спектральных линий, доступных в сети Интернет.

В первом разделе работы определены структура элемента банка данных и дополнительные атрибуты физических величин, входящих в параметры спектральной линии. Указаны ограничения при формировании банка данных.

Во втором разделе работы описываются операции над банками данных, необходимые для их формирования. Отмечается необходимость хранения в информационной вычислительной системе (ИВС) банков данных, содержащих первичные наборы данных. Описаны ограничения на первичные наборы данных и бинарные операции над ними.

В четвертом разделе описана реализация этих операций в рамках запросов SQL к параметрам спектральных линий, хранящимся в реляционной базе данных.

Разрабатываемые программные модули для работы с параметрами спектральных линий являются частью информационной системы «Атмосферная спектроскопия» (http://saga.atmos.iao.ru).

1. Описание элемента банка параметров спектральных линий

Параметры спектральных линий являются наиболее востребованными данными при расчете как спектральных функций, так и радиационных характеристик атмосферы. Параметры спектральных линий объединяются в банки данных. Структура банка данных представлена на рис. 1.

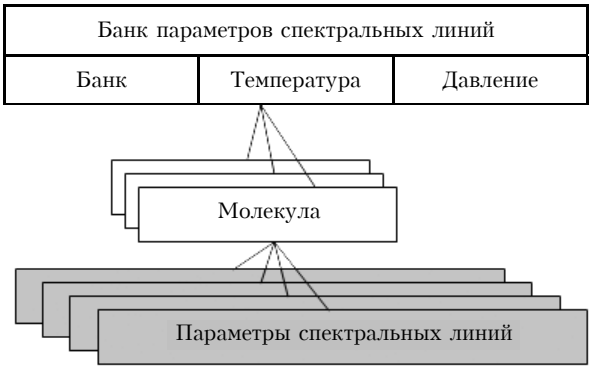


Рис. 1. Структура банка параметров спектральных линий [1]

Банки данных, например HITRAN, хранятся в файлах, содержащих набор строк одинаковой структуры. Содержанием строки являются название и код молекулы, численные значения ряда физических величин и их атрибутов, а также набор символов, идентифицирующий спектральную линию. Использование такой структуры банка данных для машинного обмена данными между информационными системами создает ряд технических проблем, которые можно избежать, выбрав новую структуру банка параметров спектральных линий.

Прежде чем описывать новый способ формирования банка данных, обратимся к детализации нижней строки рис. 1. В табл. 1 представлены физические величины и их атрибуты, входящие в число параметров спектральной линии. В ней курсивом выделены атрибуты, которыми мы предлагаем дополнить стандартный (на настоящее время) список величин и атрибутов, отображенных в таблице жирным шрифтом. Знаком «плюс» отмечено наличие у физической величины соответствующего атрибута.

Величина ошибки в формате HITRAN'a описывается номером класса. Достаточно часто в публикации используется абсолютное значение ошибки при измерении физической величины. Для учета такой возможности введен атрибут «тип ошибки», принимающий значения «абсолютный» или «интервальный». В соответствии с этим значение ошибки может быть целым или действительным числом. Для качественного понимания природы получения значения физической величины введен атрибут «тип значения», принимающий значения «эксперимент» или «расчет». Для возможности ввода данных, содержащих экспериментальные значения интенсивности линии в относительных единицах, предусмотрен атрибут «масштаб», принимающий значения «абсолютный» или «относительный».

В ИВС «Атмосферная спектроскопия» используется следующий перечень буферных газов: *воздух, водяной пар, молекулярный кислород, молекулярный азот, аргон, гелий, неон, углекислый газ*. В стандартном банке параметров спектральных линий в качестве буферного газа используется только воздух.

Таблица 1

Перечень физических величин, входящих в число параметров спектральной линии, и их атрибуты

Физическая величина	Обозначение	**Библиографическая ссылка (b)**	**Величина ошибки (e)**	*Тип ошибки* (*t*)	*Тип значения* (*v*)	*Масштаб* (*s*)
Идентификация	ID					
Частота перехода	TF	+	+	+	+	
Интенсивность линии	LI	+	+	+	+	+
Ширина, обусловленная столкновением с *буферным газом*	HW	+	+	+	+	
Ширина, обусловленная механизмами самоуширения	HWS	+	+	+	+	
Сдвиг линии давлением	SH				+	
Температурная зависимость для ширины, обусловленной столкновениями *с буферным газом*	TD				+	
Коэффициент Эйнштейна	CE					
Нижний уровень энергии	LSE					
Нижний статистический вес	LSW					
Верхний статистический вес	USW					
Флаг для смешения	FM					

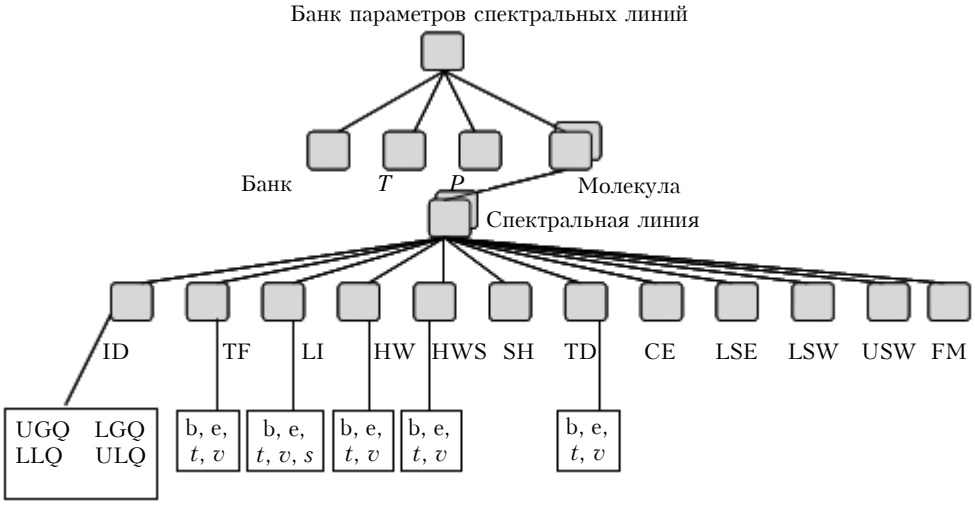


Рис. 2. Детальная структура банка параметров спектральных линий

На рис. 2 приведена детальная структура банка данных параметров спектральных линий, содержащая все дополнения, предложенные выше. Такая структура является деревом и описывается с помощью языка разметки данных XML [6]. Предложенная нами XML-схема для описания такой структуры представлена в Интернете [7].

Стоит отметить ограничение на параметры спектральных линий из банка данных. Для каждой молекулы в банке параметров спектральных линий не должно содержаться ни одной пары спектральных линий с одинаковой идентификацией. Названия банка и связанных с ним атрибутов, температура и давление относятся ко всем параметрам спектральных линий банка.

2. Операции преобразования наборов первичных данных

Набором данных (параметров спектральных линий) назовем данные, имеющие структуру банка параметров спектральных линий, но не содержащие значения температуры и давления. Первичным набором данных назовем такой набор, в котором атрибут «Библиографическая ссылка» имеет одинаковое значение для всех физических величин, входящих в набор данных, а каждый из атрибутов «тип значения», «тип ошибки» и «масштаб» имеет неизменяемое значение для конкретной физической величины. Система ввода данных в ИВС «Атмосферная спектроскопия» позволяет создавать только первичные наборы данных (банки). Операции над наборами данных (банками данных) дают пользователю возможность формировать составные банки данных. Операции со строкой или строками данных являются ключевым механизмом при формировании составных источников данных.

В данной работе к операциям преобразования наборов данных относятся те операции, которые могут изменять структуру данных, но не меняют значения физических величин и их атрибутов. Выделены два вида операций: унарные и бинарные.

К числу унарных операций относятся выборка строк по числовым отношениям (=, <, >) для физических величин (например, выборка слабых или сильных линий), выборка по символьному значению атрибута (например, все строки с экспериментальными значениями интенсивности) и выборка с помощью атрибутов, связанных с идентификацией спектральной линии (выборка линии или спектральной полосы). К этим операциям добавлена операция редукции числа физических величин в строке (за исключением идентификации, частоты перехода и интенсивности линии).

Бинарные операции со строками данных включают в себя:

1) выборки линий с одинаковой колебательно-вращательной идентификацией в паре наборов данных (**пересечение** набора данных);

2) выборки в паре наборов параметров спектральных идентифицированных линий, не имеющих пары по идентификации (**дополнение** набора данных);

3) объединение наборов данных, не содержащих строки с одинаковой идентификацией (**объединение** наборов данных);

4) объединение наборов данных, в которых для каждой идентифицированной линии одного набора параметров спектральных линий имеется соответствующая идентифицированная линия другого набора параметров спектральных линий (операции **соединения** для пары строк с одинаковой идентификацией).

Отметим, что операции соединения некоммутативны.

Результаты операций объединения и соединения для двух элементарных наборов данных, представленных на рис. 3, показаны на рис. 4 и 5. Все сокращения, используемые на рис. 3–5, описаны в табл. 1.

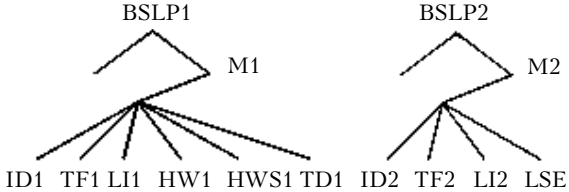
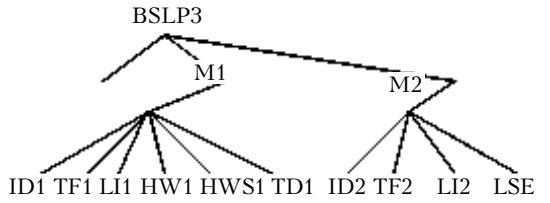
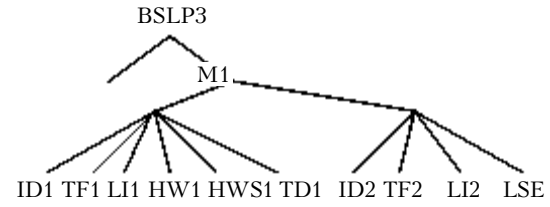


Рис. 3. Исходные элементарные наборы данных



а



б

Рис. 4. Варианты результатов применения операции объединения: *а* — M1 ≠ M2; *б* — M1 = M2, ID1 ≠ ID2

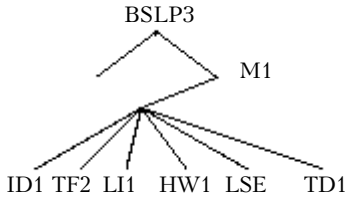


Рис. 5. Результат операции соединения: M1 = M2, ID1 = ID2

3. Хранение данных

Неудобства фиксированного формата текстового файла HITRAN'а можно устранить, записав данные с XML-разметкой [www.w3c.com]. В этом случае можно говорить не о формате файла, а о его структуре. А структура, как было сказано ранее, меняется редко и незначительно. Таким образом, изменения представления чисел или количества значащих знаков не влияют на способ работы с таким файлом. Структура файла описывается XML-схемой, ссылка на которую может быть размещена в файле.

Хранение данных в формате XML позволяет хорошо структурировать и описывать (с помощью XML-схемы) данные. Можно сказать, что XML-разметка хорошо подходит для распространения данных как альтернатива текстовому файлу с жестким форматом. Недостатком такого представления данных является то, что поиск в XML-файлах абсолютно не эффективен, так как этот формат предназначен в первую очередь для хранения данных.

Обеспечить структурированное хранение данных и хорошие поисковые возможности сегодня позволяют системы управления базами данных (СУБД). Возможности структурирования данных в СУБД соизмеримы и близки к возможностям, заложенным в XML-разметку.

СУБД достаточно хорошо подходит в качестве средства хранения и доступа к данным. С помощью СУБД можно хранить структурированные данные, осуществлять эффективные поиск, выборку и другие операции над хранимыми данными.

4. Реализация операций с данными на реляционной алгебре

Современные СУБД предлагают множество средств для работы с данными. Одним из таких средств является язык запросов SQL. Этот язык достаточно удобен, так как многие необходимые операции можно выполнить непосредственно с его помощью, не прибегая к дополнительным средствам.

Рассмотрим реализацию операций с использованием СУБД с поддержкой SQL. При описании операций будем пользоваться понятиями, используемыми при описании реляционной модели [8].

Унарные операции

Выборка определяет результирующее отношение, которое содержит только те кортежи, которые удовлетворяют заданному условию.

Проекция определяет новое отношение, содержащее вертикальное подмножество исходного отношения, создаваемое посредством извлечения значений указанных атрибутов и исключения из результата строк дубликатов.

Эти операции в полной мере реализуются оператором SELECT языка запросов SQL. Синтаксис этого оператора позволяет выполнять указанные операции как по отдельности, так и совместно.

Бинарные операции

Объединение — конкатенация всех кортежей из исходных отношений в одно отношение с удалением дублирующихся кортежей. Результирующее отношение будет иметь ту же степень, а кардинальность — равную или меньшую сумму кардинальностей исходных отношений. В случае с параметрами спектральных линий исходные отношения не должны пересекаться по спектральному диапазону и по колебательно-вращательной идентификации спектральных линий.

Эта операция выполняется с помощью операторов SELECT и UNION, осуществляющих выбор наборов из операндов и объединение их в один. Результирующий набор записывается в базу данных (БД).

Соединение — это соединение по эквивалентности, выполненное по набору атрибутов. Степень отношения будет равна сумме степеней исходных отношений минус количество атрибутов, использующихся для соединения. Возможен некоммутативный вариант соединения — левое или правое открытое соединение, когда из одного операнда выбираются все кортежи, а из другого только те, что удовлетворяют условию соединения.

В случае с параметрами спектральных линий соединение может производиться только по колебательно-вращательной идентификации спектральных линий. Соединение в описанном выше виде может быть выполнено только для просмотра данных, но не для занесения в БД, так как в БД имеется только по одному атрибуту кортежа (по одному столбцу в строке) для каждого параметра спектральной линии.

Чтобы выполнить это условие, необходимо выполнить такую проекцию над результатом соединения, чтобы в результирующих кортежах осталось по одному атрибуту, соответствующему каждому параметру линии. Если обратиться к примеру, рассмотренному ранее (см. рис. 5), то правила, по которым осуществлялась проекция, можно найти в табл. 2. Плюс в колонке таблицы означает, что значение данной физической величины присваивается соответствующей физической величине в результирующей структуре данных.

Таблица 2

Правила для формирования проекции

Набор данных	I	TF	HWS	SH	TD	LSE
1	+	+			+	
2			+	+		+

В результате ограничений на набор параметров спектральных линий получаем, что операции соединения, описанные в разд. 2, состоят из двух операций в терминах реляционной алгебры — соединения и проекции.

Для выполнения этих операций на языке SQL используется оператор SELECT совместно с оператором JOIN. В паре эти операторы позволяют получить необходимый результат.

Заключение

В работе предложено дополнение к списку физических величин, используемых в качестве параметров спектральных линий в банке данных HITRAN. Расширенную структуру данных предлагается применять для обмена данными в сети Интернет между информационными системами. Описаны операции над наборами параметров спектральных линий. Предложена реализация этих операций для данных, размещенных в реляционных базах данных.

Введенные операции после их формального описания в рамках RDF-описания могут использоваться в качестве метаданных. Такого рода метаданные характеризуют процессы формирования комплексных источников данных.

Авторы благодарны РФФИ за финансирование работы (грант № 05-07-90196).

1. *Rothman L.S., Jacquemart D., Barbe A., Chris Benner D., Birk M., Brown L.R., Carleerf M.R., Chackerian C., Chancea Jr.K., Dana V., Devi V.M., Flaud J.-M., Gamache R.R., Goldman A., Hartmann J.-M., Jucks K.W., Maki A.G., Mandin J.-Y., Massie S.T., Orphal J., Perrin A., Rinsland C.P., Smith M.A.H., Tennyson J., Tolchenov R.N., Toth R.A., Vander Auwera J., Varanasi P., Wagner G.* The HITRAN 2004 Molecular Spectroscopic Database, http://www.hitran.com
2. *Бабиков Ю.Л., Барб А., Головко В.Ф., Тютерев Вл.Г.* Интернет-коллекции по молекулярной спектроскопии // Сб. трудов 3-й Всерос. конф. по электронным библиотекам. Петрозаводск, 2001. С. 183–187. Spectroscopy of Atmosphic Gases, http://spectra.iao.ru
3. *Mikhailenko S., Babikov Yu., Tyuterev Vl.G., Barbe A.* The DataBank of Ozone Spectroscopy on WEB (S&MPO) // Computat. Technol. 2002. V. 7. (Специальный выпуск) P. 64–70. Spectroscopy and molecular properties of Ozone, http://ozone.iao.ru
4. *Быков А.Д., Воронин Б.А., Козодоев А.В., Лаврентьев Н.А., Родимова О.Б., Фазлиев А.З.* Информационная система по молекулярной спектроскопии. 1. Работа с данными // Оптика атмосф. и океана. 2004. Т. 17. № 11. С. 921–926. Атмосф. спектроскопия, http://saga.atmos.iao.ru
5. *Sowa J.F.* Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 2000. 594 p. Top-Level Categories, http://www.jfsowa.com/ontology/toplevel.htm
6. *Extensible* Markup Language (XML) 1.0 (Third Edition), http://www.w3.org/TR/2003/PER-xml-20031030
7. XML-схема для описания банка параметров спектральных линий, http://saga.atmos.iao.ru/data/xsd/SpectralLines-3.xsd
8. *Конноли Т., Бегг К., Страчан А.* Базы данных: проектирование, реализация и сопровождение. Теория и практика. 2-е изд. М.: Издательский дом «Вильямс», 2000. 1120 с.

A.V. Kozodoev, A.Z. Fazliev. **Information system for molecular spectroscopy. 2. Transformation operations over the set of the spectral line parameters.**

A line structure of data is used for storing the spectral line parameters in the HITRAN and GEISA data banks. This structure is convenient for usage in applications, but practically is not effective for storing data in an information system and is inapplicable for computer data processing. In this paper, we suggest a modified data structure of spectral line parameters. The modification includes the addition of three attributes for physical quantities and the values of buffer gas-broadened width and the temperature dependence of these widths. The list of the buffer gases consists of the water vapor, argon, neon, carbon dioxide, and so on. The modified structure of the data is used for collective processing of these data. A set of operations is proposed for the compilation of the compound data bank. The implementation of the operations is discussed for the spectral line parameters stored.