

А.В. Козодоев, А.И. Привезенцев, А.З. Фазлиев

## Информационная система для решения задач молекулярной спектроскопии. 3. Уровни энергии молекул

*Институт оптики атмосферы СО РАН, г. Томск*

Поступила в редакцию 28.05.2007 г.

Описана часть информационной системы по молекулярной спектроскопии, построенная на основе онтологии задач. Приведено описание двух задач (прямой и обратной) по нахождению уровней энергии молекул. Наполнение базы данных, связанных с этими задачами, проведено для двух трехатомных молекул: воды, сероводорода и их изотопмеров.

### Введение

Данные о спектральных параметрах молекул лежат в основе многих современных методов исследования газовых сред. Их классификацией, сбором, хранением и распространением занимаются несколько групп, среди которых выделяется группа, поддерживающая банк данных HITRAN [1]. Стоит отметить, что задача классификации данных относится к задачам спектроскопии, и ее решение во многом определяет организацию сбора, хранения и распространения данных и знаний — действия, являющиеся предметом изучения в информатике.

За почти сорокалетнюю историю группа, поддерживающая HITRAN, несколько раз модифицировала как набор физических сущностей, так и формат документов и файлов, в котором хранятся данные. В последнее десятилетие произошли события, которые заставляют критически рассматривать существующую форму сбора, хранения и распространения спектральных данных, используемых этой группой. Во-первых, появились работы по расчету вариационными методами спектра воды, содержащие сотни миллионов спектральных линий (см., например, [2, 3]). Из-за такого количества линий файловая структура стала неудобной формой для хранения данных, имеющих объем в десятки гигабайт. Во-вторых, при решении задач вариационным методом появились квантовые числа для описания уровней энергии воды, которые корректнее описывают спектр в диапазоне частот более  $20000\text{ см}^{-1}$  [3] и возникла техническая проблема сравнения больших объемов данных. В-третьих, Интернет сделал более доступными большие объемы информации как в целом, так и по частям. Все перечисленные особенности указывают на то, что настала необходимость перехода от работы с данными, хранящимися в файле, к информационно-вычислительным системам (ИВС), предоставляющим пользователям графические интерфейсы для работы с данными и информационными процессами, характерными для молекулярной спектроскопии.

Первые такого рода системы возникли семь лет тому назад [4]. Однако дальнейшее интенциональное

развитие структур данных и метаданных, в частности составление онтологий, происходило при создании ИВС «Атмосферная спектроскопия» [5–7]. Это развитие было основано на подходе семантического грида [8] к проектированию современных интеллектуальных ИВС. В рамках этого подхода ИВС представляется в виде трех слоев: слоя данных и вычислений, информационного слоя и слоя знаний. Такое разделение позволяет явно очертить области моделирования. Моделирование в предметной области связано с уровнем данных и вычислений. Моделирование в двух других слоях относится к задачам информатики и связано с информационными объектами и процессами работы с ними.

Данная статья посвящена описанию части ИВС «Атмосферная спектроскопия», содержащей результаты решения двух типов задач: прямой и обратной задач нахождения уровней энергии молекул. Цель работы заключается в создании базы данных и базы знаний для той части молекулярной спектроскопии, которая относится к уровням энергии молекул, в проверке данных на ограничения, характерные для разных типов молекул, проверке знаний на согласованность и принадлежность к каталогам, формировании метаданных и обеспечении сравнения уровней энергии из разных источников данных по значениям уровней энергии или значениям квантовых чисел. Экстенциональная часть метаданных формируется автоматически при сборе данных, конструировании составных источников данных и решении задач в ИВС. Слой знаний содержит в себе понятия, формализованные средствами семантического веба.

### 1. Данные и метаданные об уровнях энергии

Уровни энергии молекул определяются в двух задачах молекулярной спектроскопии: при решении прямой задачи на основе методов квантовой механики и при решении обратной задачи на основе подготовленного списка переходов (пиклиста), полученного из обработки результатов измерений.

Решение прямой задачи предоставляет информацию об уровнях энергии молекулы, включающую в себя данные о значениях уровней энергии и квантовых числах. Решение обратной задачи включает в себя данные о значениях уровней энергии, квантовых числах, величинах неопределенности значений уровней энергии и числе переходов, использованных для расчета уровня энергии. Перечисленные данные хранятся в базе данных информационной системы. Заметим, что рассчитанные в прямой и обратной задачах уровни энергии имеют разное происхождение, т.е. являются разными физическими сущностями.

Информация об уровнях энергии, наряду с перечисленными выше данными этих задач, включает в себя метаданные. Метаданные характеризуют уровни энергии по отношению, например, к статье, в которой они были опубликованы, по отношению к количественным характеристикам данных, полученных при измерении или расчете, по отношению к методу, использованному при расчете, по отношению к входным данным задачи и т.д. На рис. 1 приведен скриншот, представляющий метаданные об уровнях энергии, относящихся к источнику данных, сформированному на основе решения прямой задачи. В отличие от квантовых чисел и погрешностей метаданные относятся не к одному уровню энергии, а ко всему массиву уровней энергии, являющемуся компонентой источника данных. Они необходимы для описания декларативных знаний в молекулярной спектроскопии, связанных с систематизацией или, другими словами, с построением таксономий информационных ресурсов в молекулярной спектроскопии.

Структура массива данных, относящегося к прямой и обратной задачам и загружаемого пользователем в ИВС, обязательно должна содержать уровни энергии и квантовые числа. Заметим, что для молекулы воды допускается использование трех типов нотаций квантовых чисел: нормальные моды, нотации Швенке [9] и Барбера–Теннисона (BT2) [3].

В настоящее время заполнение базы данных является фрагментарным, т.е. в ней представлены уровни энергии только для молекул воды и сероводорода. Для молекулы воды расчетные данные содержат значения уровней энергии до  $30\,000\text{ см}^{-1}$ , а данные, полученные из измерений при решении обратной задачи, содержат значения уровней энергии до  $26\,000\text{ см}^{-1}$ . Верхняя граница значений полного углового момента  $J$  для расчетных данных равна 50 и для экспериментальных данных 42. Наибольшее число расчетных уровней энергии превышает 220 000, а число уровней энергии, рассчитанных из экспериментальных данных, превышает 20 000. Данные об уровнях энергии для других изотопомеров воды содержат от нескольких сотен до нескольких тысяч уровней энергии. В ИВС занесены данные Е.Р. Половцевой [10], включающие в себя более 2200 уровней энергии для сероводорода в диапазоне частот  $5800\text{--}10950\text{ см}^{-1}$  и несколько сотен уровней энергии для изотопомеров  $\text{H}_2^{34}\text{S}$  и  $\text{H}_2^{33}\text{S}$ .

База знаний содержит описание 70 источников данных в виде утверждений, сформированных средствами языка OWL. С помощью свободно распространяемого программного обеспечения Protégé [11] пользователь, импортировав базу знаний, может осуществить семантический поиск источников данных по значениям метаданных.

**Аннотация ( 2007\_Zobov\_H2\_160 )**

Входные данные		Выходные данные	
Вещество	H <sub>2</sub> O	<b>Уровни энергии</b>	
Потенциальная функция	UNDEFINED	Единица измерения	см <sup>-1</sup>
Массы атомов	UNDEFINED	Минимальное значение	23.794
Базисные волновые функции	UNDEFINED	Максимальное значение	27888.028
<b>Метод (название и ссылка)</b>		Число уровней энергии	28232
DVR3D		<b>Квантовые числа</b>	
<b>Источник данных</b>		Тип квантовых чисел	NormalModes
N.F. Zobov, R.I. Ovsyannikov, S.V. Shirin, O.L. Polyansky, Labeling of H <sub>2</sub> (16)O, H <sub>2</sub> (17)O and H <sub>2</sub> (18)O synthetic line lists calculated by variational methods in the 0 - 26000 cm <sup>-1</sup> range. // Optics and Spectroscopy, 2007, v. 102, no.3, p.394-399.		Максимальное значение для полного углового момента J	10
Quantum numbers have been assigned in the theoretical spectra of three isotopologues of the water molecule: H <sub>2</sub> <sup>16</sup> O, H <sub>2</sub> <sup>17</sup> O, and H <sub>2</sub> <sup>18</sup> O. The spectra were calculated by variational methods in the region 0–26 000 cm <sup>-1</sup> at a temperature of 296 K. For each molecule, the quantum numbers are assigned to more than 28 000 levels. The quantum numbers are assigned to 216 766, 211 679, and 211 073 spectral lines of the H <sub>2</sub> <sup>16</sup> O, H <sub>2</sub> <sup>17</sup> O, H <sub>2</sub> <sup>18</sup> O molecules, respectively. The theoretical spectra with the assigned quantum numbers are available in the Internet.		Число уровней энергии с уникальной идентификацией	28231
		Число уровней энергии с не уникальной идентификацией	1
		<b>Квантовые числа</b>	
		Тип квантовых чисел	BT2
		Максимальное значение для полного углового момента J	10
		Число уровней энергии с уникальной идентификацией	28232
		Число уровней энергии с не уникальной идентификацией	0

Рис. 1. Метаданные источника данных, содержащего уровни основного изотопомера энергии воды

## 2. Сбор и хранение данных об уровнях энергии

Пользователю ИВС предоставлена возможность загрузки собственных данных об уровнях энергии в базу данных и базу знаний с целью сравнения их с данными, рекомендованными экспертами. Для этой цели разработано и реализовано программное обеспечение, позволяющее на основе XML-схем контролировать соответствие входных данных типовым ограничениям, характерным для данных об уровнях энергии. Для классификации данных молекулярной спектроскопии нами использована модель, в которой предметная область представлена в виде иерархии задач [12].

Ключевым концептом для формирования ИВС на основе разбиения данных по задачам молекулярной спектроскопии является *источник данных*. Источники данных могут быть элементарными и составными. Элементарным источником данных является источник, в котором содержатся данные об уровнях энергии, опубликованные в одной статье и относящиеся к молекуле одного типа. Все прочие источники данных являются составными, например, HITRAN и Geisa.

Для элементарного или составного источника данных используются четыре уровня описания.

1. *Описание атрибутов источника данных*: название источника данных, ссылка на публикацию, комментарий, тип задачи и название вещества, уровни энергии которого относятся к источнику данных.

2. *Описание интенционала данных, представленных в источнике данных*: уровни энергии, квантовые числа, погрешности и число переходов, использованных для получения уровня энергии.

3. *Метаданные о задаче, связанной с источником данных*, определяются в онтологиях задач спектроскопии и содержат данные о входных данных и методе решения задачи. В частности, для прямой задачи нахождения уровней энергии воды ими являются URI информационных ресурсов, таких как потенциальная функция и базисные волновые функции.

4. *Метаданные об уровнях энергии* содержат данные о минимальном и максимальном значениях уровня энергии, числе уровней энергии, типе квантовых чисел, максимальном значении полного углового момента, числе уникальных и не уникальных значений квантовых чисел, приписанных уровням энергии.

Для сбора массивов данных, входящих в элементарные источники данных, создана система загрузки с веб-интерфейсом. Загрузка уровней энергии, полученных при решении как прямой, так обратной задач, состоит из трех этапов:

1. Выбор источника данных, задание молекулы и числа физических сущностей в загружаемом массиве данных.

2. Формирование упорядоченного списка физических сущностей в порядке, определяемом следованием колонок в загружаемом файле (рис. 2).

Колонки файла

№	Физическая величина	Начальная позиция колонки	Конечная позиция колонки
1	Квантовые числа. Нормальные моды	1	24
2	Уровень энергии (см <sup>-1</sup> )	26	36
3	Квантовые числа. BT2	38	50
4	Квантовые числа. Швенке	52	76

Ok

Рис. 2. Описание колонок файла, содержащего уровни энергии и квантовые числа

### 3. Задание адреса нахождения файла с данными.

Пользователь ИВС может самостоятельно формировать собственные источники данных. Для связи источника данных с библиографической ссылкой используется база данных, включающая в себя сотни библиографических ссылок в области молекулярной спектроскопии. При отсутствии необходимой библиографической ссылки пользователь может формировать собственные ссылки на информационные ресурсы. Сформированные пользователем библиографические ссылки не будут доступны другим пользователям.

Система позволяет определять следующие ошибки формирования файла пользователя: несоответствие типов данных, использованных пользователем в файле; некорректность структуры физических сущностей, описанной пользователем при подготовке загрузки (второй шаг); несоответствие допустимым интервалам значений для физических величин и т.д.

В настоящее время в информационной системе собрано 70 элементарных источников данных об уровнях энергии молекул H<sub>2</sub>O (H<sub>2</sub><sup>17</sup>O, H<sub>2</sub><sup>18</sup>O, HDO, HD<sup>18</sup>O, D<sub>2</sub>O), H<sub>2</sub>S (H<sub>2</sub><sup>34</sup>S, H<sub>2</sub><sup>33</sup>S). Большая часть источников данных связана с решением обратных задач. Отдельные источники данных разделены на части, содержащие уровни энергии *ortho*- и *para*-воды. Наиболее полно представлены данные об уровнях энергии молекулы воды. Отметим, что в ИВС можно загружать данные об уровнях энергии молекул SO<sub>2</sub> и O<sub>3</sub>, обладающих тем же набором квантовых чисел, что и вода.

На рис. 3 показана упрощенная схема базы данных, представляющая связи таблиц, в которой хранятся уровни энергии. Упрощение касается перечня используемых нотаций квантовых чисел и связи с библиографической базой данных.

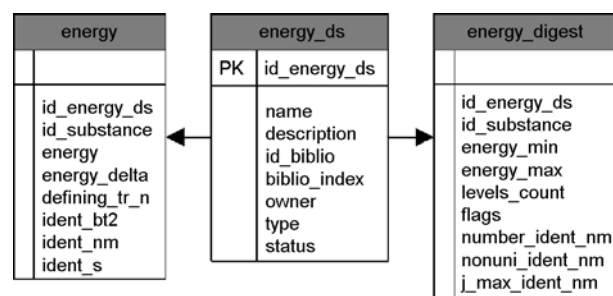


Рис. 3. Упрощенная структура базы данных уровней энергии

### 3. Поиск и сравнение уровней энергий разных источников данных

Уровни энергии, связанные с источниками данных и относящиеся к прямой или обратной задачам, можно сравнивать между собой как по значению уровня энергии, так и по квантовым числам. Такое сравнение важно при анализе экспериментальных данных, число которых на порядок меньше расчетных данных. Особое значение оно принимает для уровней энергии, превышающих  $20\,000\text{ см}^{-1}$ .

Пользователь может проводить сравнение по значению уровней энергии с заданной точностью от 2 до 7 значащих цифр (единицы измерения  $\text{см}^{-1}$ ). Выбор ограничений на значения полного углового момента  $J$  позволяет выделять в сравниваемых источниках данных массивы уровней энергии с  $J$  в заданном пользователем интервале, что позволяет отсекаать, например, в расчетных данных уровни, которые пока еще нельзя восстановить из измерений. Сравнение можно проводить как в табличном, так и графическом виде. Скриншот табличного сравнения показан на рис. 4.

Сравнение уровней энергии молекул по квантовым числам (нормальные моды) позволяет выделить уровни энергии с совпадающими квантовыми числами и предоставить разницу в значении соответствующих уровней энергии. Для молекулы воды можно проводить сравнение источников данных, содержащих наборы квантовых чисел разной нотации. При таком сравнении квантовые числа ( $\nu_1, \nu_2, \nu_3, J, K_a, K_c$ ) преобразуются в пару квантовых чисел ( $J, s$ ), в которой квантовое число  $s$  описывает симметрию [3]. Хотя такое сравнение не дает однозначной связи между уровнями, оно является по-

лезным для сравнения значений уровней энергии, полученных в прямой и обратной задачах.

Пользователь может экспортировать данные об уровнях энергии молекул при просмотре или сравнении источников данных, используя для этой цели их табличное представление. Доставка данных осуществляется по сети Интернет в текстовом формате или в формате данных, используемых программой обработки таблиц Excel.

### Заключение

Основная идея нового подхода к сбору, хранению и классификации данных об уровнях энергии молекул воды, сероводорода и некоторых других молекул состоит в том, что уровни энергии и сопутствующие данные, загружаемые в систему, связываются с источниками данных, представляющими их при работе в ИВС. Источники данных ассоциированы с типовыми задачами молекулярной спектроскопии [12]. Наполнение базы данных расчетными и экспериментальными данными позволяет рассматривать ее в настоящее время как прототип информационной системы, ориентированной на представление данных об уровнях энергии атмосферных молекул.

Авторы благодарны J. Tennyson и A. Császár за обсуждение вариантов классификации спектральных данных, Р.Н. Толченову за консультации при организации сравнения данных, имеющих разные нотации квантовых чисел, Н.Ф. Зобову и С.В. Ширину за предоставленные результаты вычислений уровней энергии молекулы воды, А.Ю. Ахлестину за реализацию сравнения уровней энергии, а также РФФИ за финансирование работы (гранты № 06-07-89201, 05-07-90196).

Источник данных слева	Источник данных справа
1997JMS_Polyansky	2007_Zobov_H2_160
O.L.Polyansky, N.F.Zobov, S.Vitia, J.Tennyson, P.F.Bernath and L.Wallace, High-Temperature Rotational Transitions of Water in Sunspot and Laboratory Spectra. // Journal of Molecular Spectroscopy, 1997, T. 186, B. 2, C. 422-447.	N.F. Zobov, R.I. Ovsyannikov, S.V. Shirin, O.L. Polyansky, Labeling of H2(16)O, H2(17)O and H2(18)O synthetic line lists calculated by variational methods in the 0 - 26000 cm-1 range. // Optics and Spectroscopy, 2007, v. 102, no.3, p.394-399.

Число совпадающих значений уровней энергии равно 11. Округление до 3 знаков.

Показать 40		строк от 0		Всего строк 32263		> >>		Настройки										
$\nu_1$	$\nu_2$	$\nu_3$	$J$	$K_a$	$K_c$	Уровни энергии (E (см <sup>-1</sup> ))			$J$	$s$	$\nu_1$	$\nu_2$	$\nu_3$	$J$	$K_a$	$K_c$		
						212.156				3	3	2	0	0	0	3	2	1
0	0	0	4	0	4	222.053				4	1	1	0	0	0	4	0	4
						224.839				4	4	1	0	0	0	4	1	4
						275.497				4	2	1	0	0	0	4	1	3
						285.218				3	2	2	0	0	0	3	3	1
						285.417				3	4	2	0	0	0	3	3	0
0	0	0	4	2	3	300.362				4	3	1	0	0	0	4	2	3
0	0	0	4	2	2	315.779				4	1	2	0	0	0	4	2	2
0	0	0	5	0	5	325.348												
						325.349				5	3	1	0	0	0	5	0	5
0	0	0	5	1	5	326.625												

Рис. 4. Сравнение уровней энергии воды с точностью до трех знаков после запятой (у совпадающих значений энергии квантовые числа находятся в одной и той же строке)

1. Rothman L.S., Jacquemart D., Barbe A., Benner D. Chris, Birk M., Brown L.R., Carleer M.R., Chakkerian C., Jr., Chance K., Coudert L.H., 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., Auwera J. Vander, Varanasi P., Wagner G. The HITRAN 2004 molecular spectroscopic database // J. Quant. Spectrosc. & Radiat. Transfer. 2005. V. 96. N 2. P. 139–204.
2. Тютчев Вл.Г. Глобальные вариационные и эффективные методы расчетов положений и интенсивностей спектральных линий трехатомных молекул: некоторые тенденции и особенности нового поколения спектроскопических информационных систем // Оптика атмосф. и океана. 2003. Т. 16. № 3. С. 245–255.
3. Barber R.J., Tennyson J., Harris G.J., Tolchenov R.N. A high accuracy computed water line list-BT2 // Mon. Notic. Roy. Astron. Soc. 2006. V. 368. P. 1087–1094.
4. Бабиков Ю.Л., Барб А., Головки В.Ф., Михайленко С.Н., Тютчев Вл.Г. Интернет-коллекции по молекулярной спектроскопии // Сб. тр. 3-й Всерос. конф. «Электронные библиотеки: перспективные методы и технологии...». Петрозаводск, 2001. С. 183–187. Spectroscopy of atmospheric gases, <http://spectra.iao.ru>
5. Быков А.Д., Воронин Б.А., Козодоев А.В., Лаврентьев Н.А., Родимова О.Б., Фазлиев А.З. Информационная система для решения задач молекулярной спектроскопии. 1. Структура информационных ресурсов // Оптика атмосф. и океана, 2004. Т. 17. № 11. С. 921–926.
6. Козодоев А.В., Фазлиев А.З. Информационная система для решения задач молекулярной спектроскопии. 2. Операции преобразования наборов параметров спектральных линий // Оптика атмосф. и океана. 2005. Т. 18. № 9. С. 760–764.
7. Козодоев А.В., Привезенцев А.И., Фазлиев А.З. Организация информационных ресурсов в распределенной информационно-вычислительной системе, ориентированной на решение задач молекулярной спектроскопии // Вычислит. технол. Спец. выпуск. 2005. Т. 10. С. 82–91.
8. De Roure D., Jennings N., Shadbolt N. A Future e-Science Infrastructure, Report commissioned for EPSRC/DTI Core e-Sci. Program. 2001. 78 p.
9. Partridge H., Schwenke D.W. The determination of an accurate isotope dependent potential energy surface for water from extensive *ab initio* calculations // J. Chem. Phys. 1997. V. 106. N 11. P. 4618–4639.
10. Половцева Е.Р. Анализ колебательно-вращательного спектра сероводорода в области от 4500 см<sup>-1</sup> до 11000 см<sup>-1</sup>: Автореф. дис. ... канд. физ.-мат. наук. Томск, 2006. 22 с.
11. Protégé-ontology editor and knowledge-base framework, <http://protege.stanford.edu/>
12. Bykov A.D., Kozodoev A.V., Privezentsev A.I., Sinitza L.N., Tonkov M.V., Filippov N.N., Fazliev A.Z., Tretyakov M.Yu. Distributed information system on molecular spectroscopy // Proc. SPIE. 2006. V. 6580. P. 65800W.

A.V. Kozodoev, A.I. Privezentsev, A.Z. Fazliev. **Information system for molecular spectroscopy.**

### 3. Molecule energy levels.

A part of the molecular spectroscopy system is described based on the problem ontology. Two problems (direct and reverse) of finding molecule energy levels are considered. Populating of databases concerned with these problems was carried out for two triatomic molecules, i.e. water and hydrogen sulfide, and their isotopes.