ИСПОЛЬЗОВАНИЕ ТАБЛИЧНОГО ПРЕДСТАВЛЕНИЯ ПАРНЫХ ОТНОШЕНИЙ ДЛЯ СОГЛАСОВАНИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ В СПЕКТРОСКОПИИ

Ахлёстин А.Ю., Привезенцев А.И., Фазлиев А.З.
Институт оптики атмосферы СО РАН, Томск
lexa@iao.ru, remake@iao.ru, faz@iao.ru

Аннотация

В докладе описано табличное представление связанных между собой частей научных публикаций для предметной области, в которой фактологическая часть существенно превосходит понятийную. Оно позволяет давать качественную оценку состоятельности решений задач спектроскопии в случае анализа небольшого числа (до нескольких десятков) источников данных.

1. Введение

На практике системы поиска ресурсов в научных цифровых библиотеках опираются на тексты статей на естественном языке. Большая часть поисковых систем не использует формализованные понятия предметных областей, содержащиеся в искомых ресурсах в явном или неявном виде, по простой причине: большинство понятий предметной области не формализовано. В предметных областях, обладающих значительным объемом фактов (так называемых data-driven domain) можно провести формализацию. В докладе рассматривается пример такой формализации в количественной спектроскопии. Формализация количественной спектроскопии, сделанная нами, ориентирована на изучение свойств источников данных в предметной области. Анализ качества данных по спектроскопии и связанные с ним свойства источников данных (как собственные свойства источников, так и их парные отношения с другими источниками) представляют отдельную задачу. Часть этой задачи состоит в представлении парных отношений между источниками данных.

В публикациях были выделены части, содержащие решения одной из шести задач количественной спектроскопии, и которые импортированы в базы данных, каждая из которых относилась к молекуле определенной симметрии и одному из типов задач спектроскопии. Созданные базы данных содержат все опубликованные решения задач в рамках моделей данных. Каждое такое решение является частью публикации, а в количественной спектроскопии основной частью, так как содержит наибольшее число типизованных фактов. В простом случае выбирая модель публикации или ее части, можно ограничиться решением задач спектроскопии, что соответствует публикации таких решений на сайтах и FTP серверах. Решение задачи, дополненное свойствами, может служить более точной формальной моделью публикации. В спектроскопии такой набор

свойств был предложен в [1]. Наиболее полные наборы данных, присутствующие в информационной системе, относятся к молекулам воды и сероводорода [2-5].

2. Модели атомарных частей публикации

Создание моделей атомарных частей публикации для цифровой библиотеки частей научных статей связано с задачей автоматической каталогизации информационных ресурсов по количественной спектроскопии. Имеющаяся у авторов коллекция статей превышает 8000 публикаций, относящихся к периоду с 1926 года по настоящее время. Выбранная модель данных предметной области [6] характеризует решения шести задач спектроскопии, имеющих определяющее значение для прикладных предметных областей, таких как астрономия, оптика атмосферы, спектроскопия и т.д.

Основой для построения модели является допущение о том, что публикация содержит факты, являющиеся решениями ряда задач спектроскопии. Эти факты разделены на две группы. К первой группе относятся такие решения задач, которые можно отнести к одной молекуле и одному методу решения. Ко второй группе относятся все оставшиеся решения.

Извлеченные из публикаций атомарные части статей представляются в цифровой библиотеке в форме первичных или составных источников данных. Каждый источник информации представляет собой источник данных и высказывания об этом источнике данных.

Для формирования источника данных используется реляционная модель данных, а источника информации – язык онтологий OWL 2. Представление источников информации в виде индивидов онтологии является основой для автоматической каталогизации решений задач спектроскопии.

2.1. Независимые части публикации (первичные источники данных)

Разнообразие молекул, для которых решались задачи, выделенные в работе [5], и методов, которыми они решались, достаточно большое. По этой причине в одной публикации могут быть приведены решения нескольких задач разными методами и для разных молекул или их изотопологов. При систематизации данных, извлеченных из публикаций, такое смешение создает много проблем. По этой причине в работе используется информационный объект, представляющий оригинальные данные публикации, относящиеся к одной молекуле, одной задаче спектроскопии и одному методу решения. Такой объект представляет собой атомарную часть публикации, имеющую самостоятельное значение. Выбор атомарных частей публикаций в разных предметных областях может быть разным и определяется решаемыми информационными задачами.

Определение 1. Все части опубликованного решения задачи количественной спектроскопии, дополненные названием молекулы, библиографической ссылкой (или URI) и названием метода решения задачи (или URI на описание метода) называются *первичным источником данных*.

Мы предполагаем, что пустые решения не публикуются. С другой стороны решения задач могут содержать данные измерений, которые со временем устаревают или неверные решения. Количество таких источников в современной спектроскопии незначительно.

Формализованный первичный источник данных содержит решение задачи и обладает свойствами [1] (*isSolutionOf, hasMethod, isRelatedToSubstance* и *hasReference*), имеющими кардинальность равную 1. Важной характеристикой источника данных является независимость значений этих свойств от времени. Ключевым свойством в определении источника данных является *hasReference*.

В количественной спектроскопии, наряду с журналами, монографиями, отчетами и трудами конференций, в последнее десятилетие появились решения задач, размещенные в сети Интернет. Необходимость такого типа публикации обусловлена значительными их объемами (превышающими сотни Гб.)

Первичные источники данных, относящиеся к одной публикации, не имеют общих данных.

2.2. Составные источники данных (агрегации первичных данных в статьях)

Определение 2. Информационный объект, обладающий базовыми свойствами первичного источника данных, кардинальность любого из которых отличается от единицы, называется *составным источником данных*.

Примером составного источника данных является любой экспертный массив спектральных данных (например, GEISA [7]).

2.3 Источник информации

Первичный источник данных можно наделять дополнительными свойствами. Перечень и число этих свойств зависит от информационных задач, для решения которых используются такие свойства. Источник данных с дополнительными свойствами назовем источником информации.

Определение 3. Первичный источник данных, наделенный дополнительными свойствами, называется *первичным источником информации,* извлеченным из информационного ресурса (в частности, из научной статьи).

Источник информации представляет собой набор свойств и их значений, относящихся к источнику данных. Для ряда информационных задач, например, задачи поиска достоверных решений задач количественной спектроскопии, можно выбрать свойства, значения которых вычисляются автоматически. Как правило, источник информации включает в себя некоторые высказывания из публикации. Большая часть источника информации характеризует знания, содержащиеся в публикации в неявном виде.

3. Пример цифровой библиотеки атомарных частей статей

Примером цифровой библиотеки, использующей модель публикации, описанную выше, является информационная система W@DIS [8]. Она основана на коллекции публикаций по количественной спектроскопии.

Для решения ряда задач предметных областей, связанных со спектроскопией, пользователям необходима только часть фактов, содержащихся в публикациях этой коллекции. Эти факты относятся к решениям шести задач спектроскопии, связанных с нахождением параметров состояний и переходов молекул [9].

Оцифрованные факты из публикаций импортированы в информационную систему и представляют собой разные типы источников данных. При импорте данных в систему для каждого источника данных автоматически создаются высказывания об источнике данных [1], содержащий описание свойств импортированных решений задач спектроскопии.

4. Парные отношения в количественной спектроскопии

Значительную долю данных, связанных с решениями задач количественной спектроскопии, занимают значения бинарных отношений между источниками данных. В системе W@DIS использованы только три вида бинарных отношений, связанных с максимальной разностью вакуумных волновых чисел, среднеквадратическими отклонениями решений задач спектроскопии (на примере задач T2 и T6) и сравнением упорядочений квантовых чисел по связанным с ними значениями волновых чисел. В работе [10] с рядом деталей описаны как максимальная разность волновых чисел, так и среднеквадратические отклонениями решений задач спектроскопии. Ниже остановимся на сравнении упорядочений.

4.1. Сравнение упорядочений

При анализе решений задач спектроскопии можно добиться требуемой малости значений разности волновых чисел идентичных переходов и требуемых значений среднеквадратических отклонений. Однако этого бывает недостаточно для согласования решений задач спектроскопии (T2 и T6).

Пусть есть два набора A_1 и A_2 вакуумных волновых чисел и соответствующих им квантовых чисел. При этом положим, что каждому набору квантовых чисел qn_i соответствует тождественный ему набор квантовых чисел qn_j, другими словами в наборах A_1 и A_2 существуют попарно идентичные переходы. Упорядочим последовательности значений вакуумных волновых чисел по возрастанию. Припишем порядок возрастания, относящийся к волновым числам, к квантовым числам. В общем случае порядок следования квантовых чисел в двух наборах может быть различным.

Количественная оценка разупорядочения может определяться разными способами. В данной работе при описании табличного отображения разупорядочения используется следующие алгоритмы.

Коррелирующими источниками информации являются те пары источников, у которых полностью совпадают квантовые числа хотя бы одного перехода в любой нотации квантовых чисел. Для двух таких источников информации можно подсчитать различные параметры их схожести, выраженные целым числом. В нашей работе вычисляются три таких параметра. Параметром схожести (A00) назовем число перестановок переходов для совпадения порядка квантовых чисел. Параметрами схожести (A01) и (A10) назовем число исключённых переходов из первого источника информации второго источника информации для совпадения порядка квантовых чисел и наоборот, соответственно.

Число перестановок переходов для совпадения порядка квантовых чисел (A00) можно подсчитать по следующему алгоритму.

Пусть есть два массива данных X[волновое число, квантовые числа] и Y[волновое число, квантовые числа], относящиеся соответственно к первому и второму источникам

информации из корреляционной пары. Объединенный массив будет иметь вид XY[волновое число из X, волновое число из Y, совпадающие квантовые числа].

Шаг 1. Делаем сортировку массива XY по физической величине из X (вакуумное волновое число из X), получаем массив XY[волновое число из X, волновое число из Y, совпадающие квантовые числа, порядковое число при сортировке по X (id_X)].

Шаг 2. Делаем сортировку массива XY по физической величине из Y (вакуумное волновое число из Y), получаем массив XY[волновое число из X, волновое число из Y, совпадающие квантовые числа, порядковое число при сортировке по X (id_X), порядковое число при сортировке по Y (id_Y)].

Шаг 3. A00 = Сумма по всему массиву XY чисел полученных из арифметической операции над id_X и id_Y, так если (id_X > id_Y), то (id_X - id_Y-1), иначе (id_Y - id_X).

Число исключённых переходов из первого источника информации второго источника информации для совпадения порядка квантовых чисел (A01) можно подсчитать по следующему алгоритму.

Пример алгоритма приведён на языке Python.

```
#Словарь отсортированный по физической величине из массива X
rowsX = {qnsX1:wX1, ... , qnsXn:wXn}
#Словарь отсортированный по физической величине из массива Y
rowsY = {qnsY1:wY1, ... , qnsYn:wYn}
#Словарь исключённых переходов из первого источника информации второго источника
D = {}
for qnsX in rowsX:
    qnsY = reset(rowsY)
    while(True or not EOF):
        if qnsX == qnsY:
            break;
        if qnsX != qnsY:
            x = rowsX[qnsY]
            y = rowsY[qnsY]
            D[qnsY] = (x, y)
            del rowsX[qnsY]
            del rowsY[qnsY]
            qnsY = next(rowsY)
X01 =len(D)
```

Пример работы алгоритма для вычисления A01:

X = qns1:wX1 < qns2:wX2 < qns3:wX3 < qns4:wX4 < qns5:wX5 < qns6:wX6

Y = qns1:wY1 < qns5:wY2 < qns3:wY3 < qns4:wY4 < qns6:wY5 < qns2:wY6

Число исключённых переходов из первого источника информации второго источника информации A01 = 4, при D = {qns5:(wX5, wY2), qns3: (wX3, wY3), qns4:(wX4, wY4), qns6:(wX6, wY5)}

Число исключённых переходов из второго источника информации первого источника информации для совпадения порядка квантовых чисел (A10) можно подсчитать по приведённому выше алгоритму, если поменять местами массивы X и Y.

Пример работы алгоритма для вычисления A10:

X = qns1:wX1 < qns2:wX2 < qns3:wX3 < qns4:wX4 < qns5:wX5 < qns6:wX6

Y = qns1:wY1 < qns5:wY2 < qns3:wY3 < qns4:wY4 < qns6:wY5 < qns2:wY6

Число исключённых переходов из второго источника информации первого источника информации A10 = 3, при D = {qns2:(wY6, wX2), qns3: (wY3, wX3), qns4:(wY4, wX4)}

5. Табличное представление парных отношений источников данных

Импорт атомарных частей публикации в информационную систему по спектроскопии сопровождается автоматической генерацией значений свойств этих данных и формированием прикладных онтологий. Значения этих свойств хранятся в реляционной базе данных в соответствующих таблицах.

Для количественного анализа данных, относящихся к парным отношениям между источниками данных, можно применять табличное представление. Подобное представление показано на рис. 1 - 5. Оно удобно при детальном анализе качества данных и на этапе выбраковки некачественных данных. Однако такое представление требует значительного времени для анализа. Решение проблемы анализа большого числа источников информации связано с графическим представлением парных отношений.

Рис. 2-4 относятся к молекуле сероводорода и содержат количественные характеристики соответствующих бинарных отношений.

На рис.2-4 приводятся результаты значений бинарных отношений 27 источника данных, содержащих переходы для которых можно найти идентичные в других источниках данных. В таблице приведены аббревиатуры публикаций, и в соответствующей ячейке таблицы приведено отношение двух чисел. Числитель отношения указывает значение бинарного отношения, а знаменатель – число идентичных переходов в этих источниках данных.

Желтая ячейка таблицы, содержащая отношение, означает, что соответствие между идентичными переходами удовлетворяет заданному критерию, а красное – не удовлетворяет критерию.

Из анализа таблиц на рис. 2-4 следует, что сравниваемые данные могут удовлетворять одним критериям, но не удовлетворять другим. В таблицах в явном виде указаны количественные значения парных отношений. При числе источников данных, превышающем сотню, представление таблиц становится неэффективным, т.к. отсутствует возможность компактного просмотра данных на экране монитора.

Для решения некоторых задач анализа качества данных можно ограничиться качественным просмотром данных, приведенных в таблице. В этом случае из таблиц удаляются количественные характеристики, и вводят цветовое разделение на пары, удовлетворяющие количественным критериям и неудовлетворяющие им. Представленный на рис.5 интерфейс содержит данные по молекуле воды ($H_2{}^{16}O$) [10].

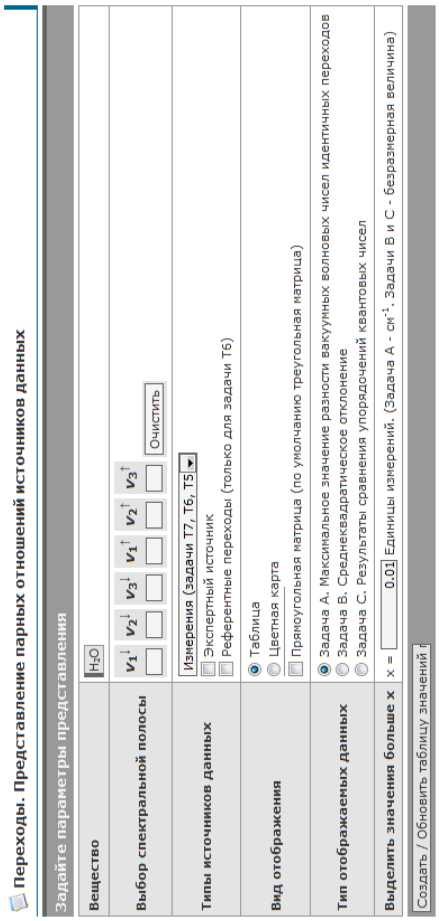

Рис.1. Значения свойств, характеризующих пары первичных экспериментальных источников данных по молекуле сероводорода.

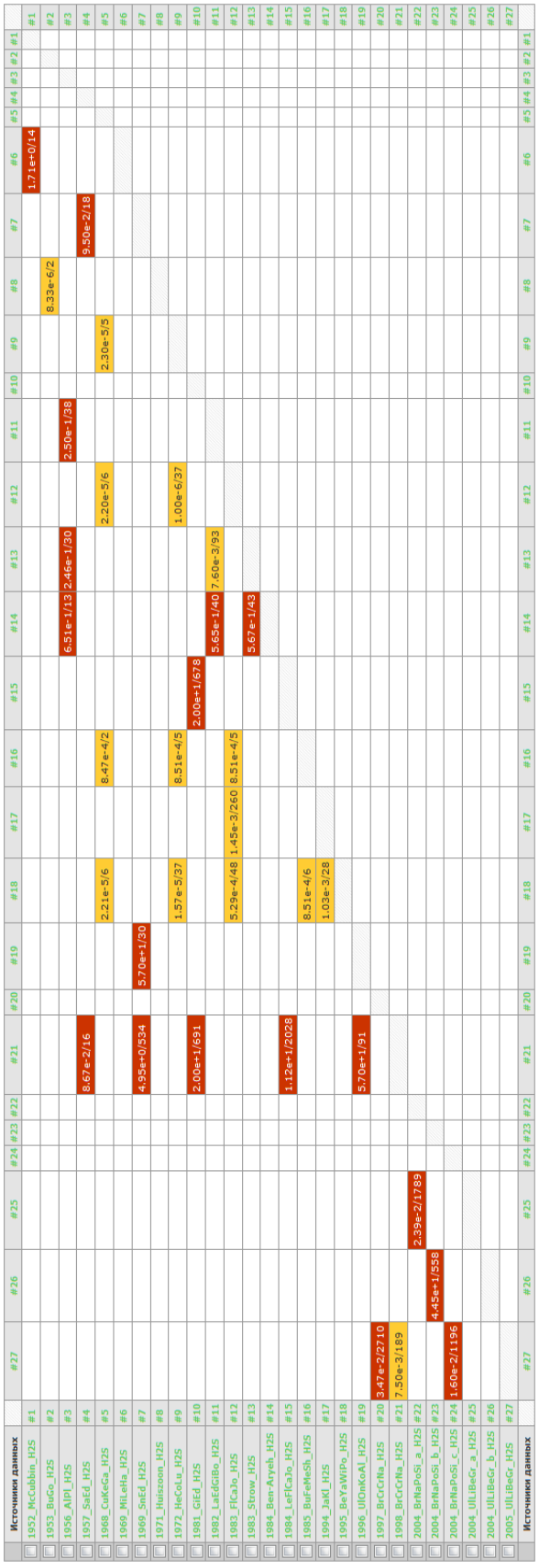

Рис.2. Значения свойств, характеризующих максимальное значение разности вакуумных волновых чисел идентичных переходов.

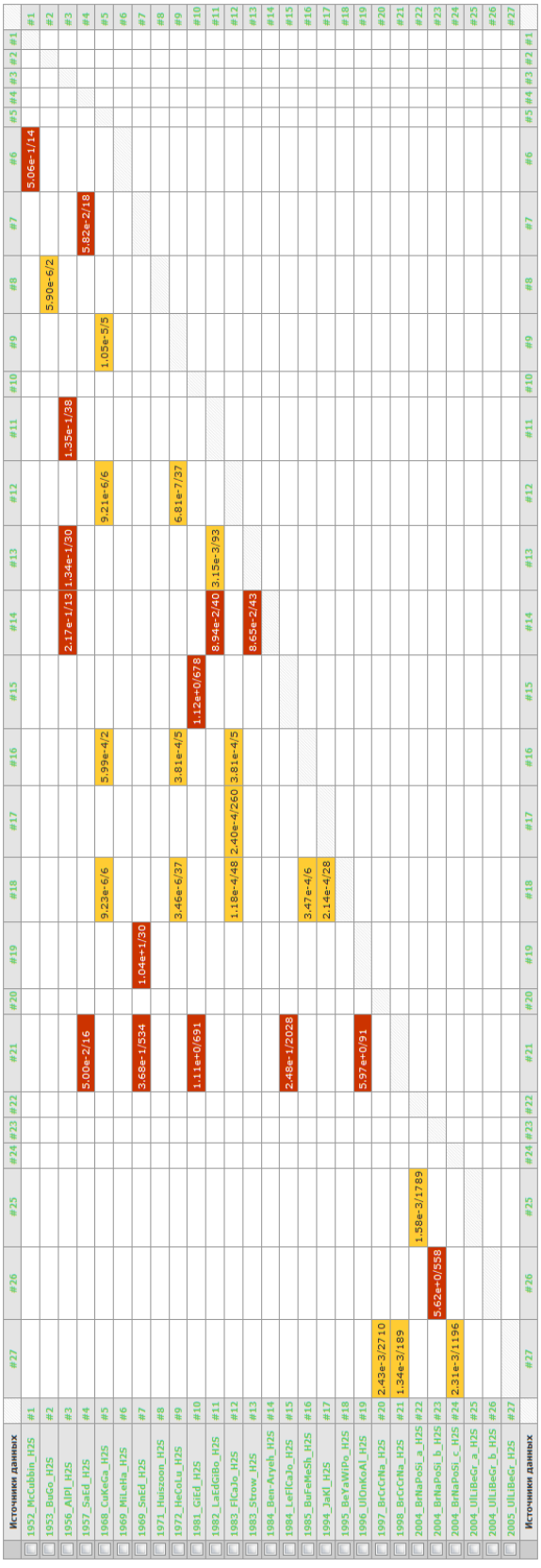

Рис.3. Значения свойств, характеризующих среднеквадратическое отклонение.

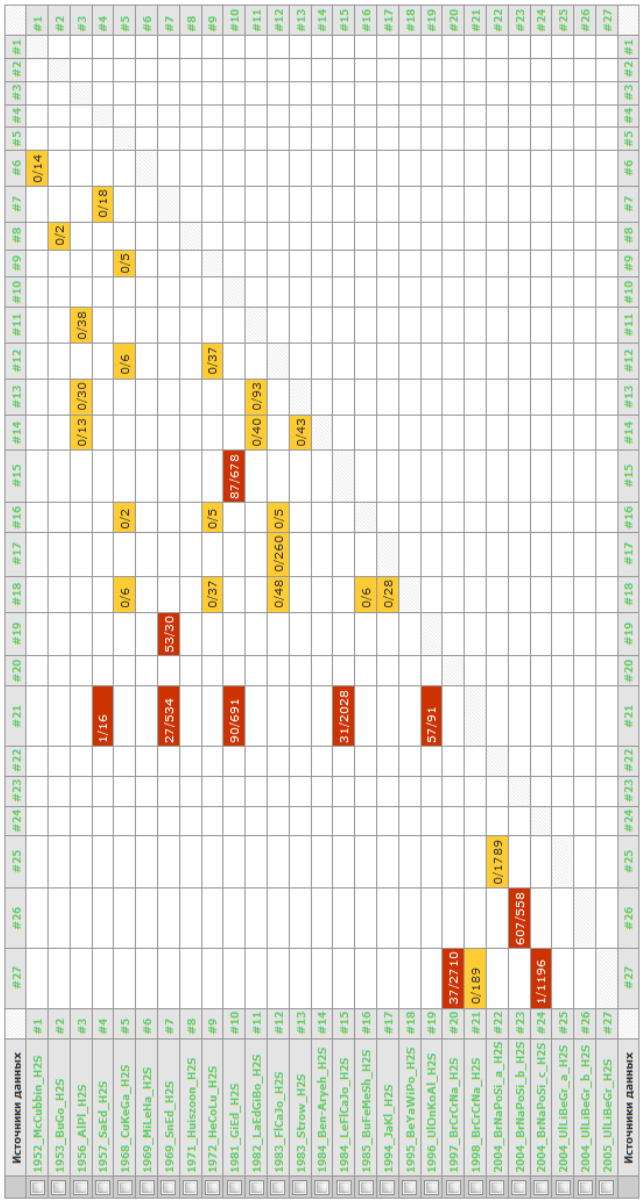

Рис.4. Значения свойств, характеризующих результаты сравнения упорядочений (A00).

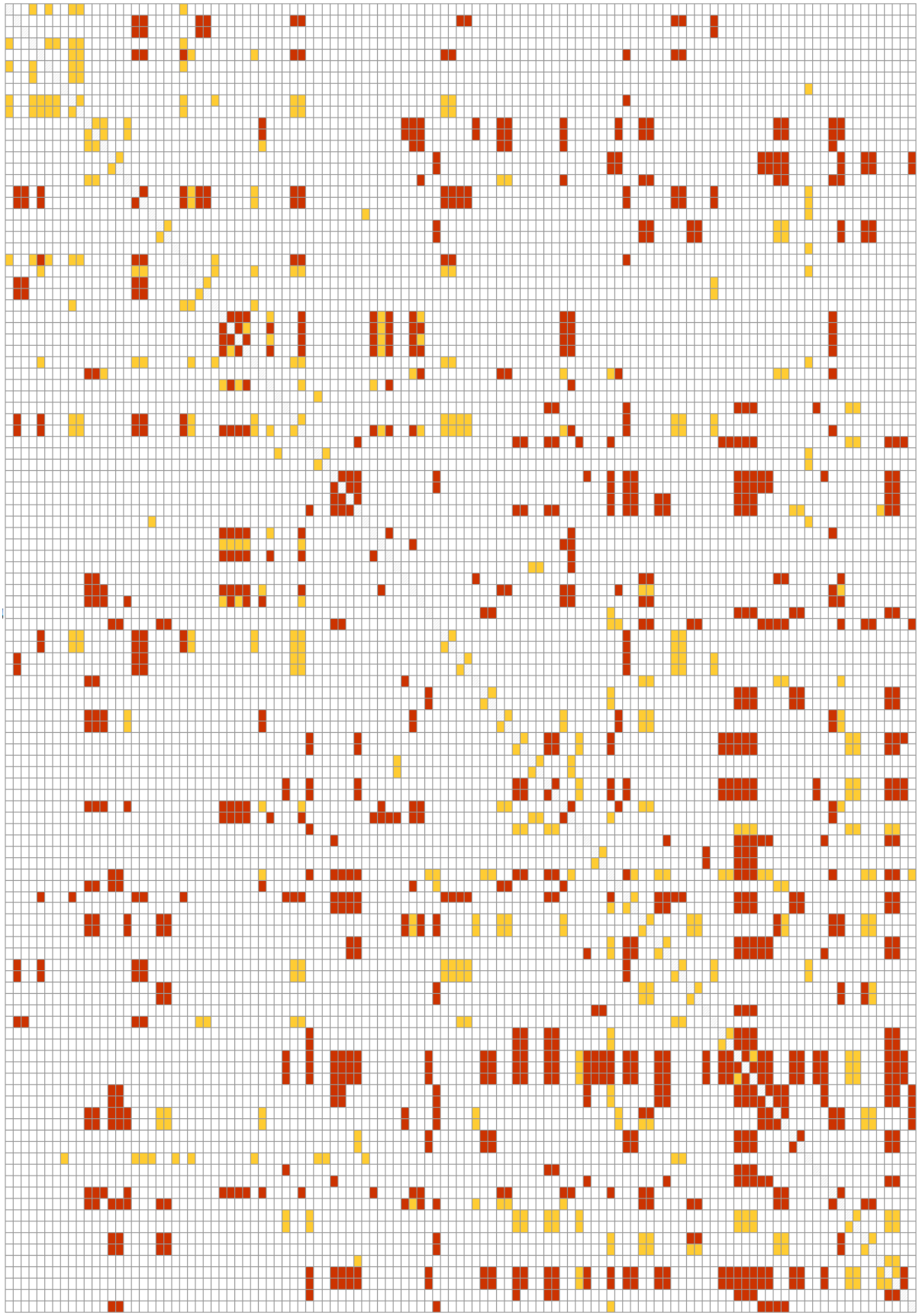

Рис.5. Цветная карта для анализа максимальных разностей волновых чисел идентичных переходов для первичных экспериментальных источников информации по молекуле $H_2{}^{16}O$. (Обозначения: ▭ - максимальная разность волновых чисел меньше 0.01 $см^{-1}$, ▬ - максимальная разность волновых чисел больше 0.01 $см^{-1}$).

6. Заключение

В работе уточнены определения источника данных и источника информации, предложенные ранее в [10], описаны параметры разупорядочения, определяющие количественную сторону оценки разупорядочения множеств, содержащих идентичные переходы. На примерах показано применение табличного представления данных для фиксации противоречий возникающих при анализе идентичных переходов с использованием разных парных отношений.

Применение описанного табличного представления может быть эффективным при проверке качества данных в виртуальном центре атомарных и молекулярных данных [11].

Авторы благодарны РФФИ (грант 11-07-00660) за финансовую поддержку.

Литература

[1] Привезенцев А.И., Царьков Д.В., Фазлиев А.З. Базы знаний для описания информационных ресурсов в молекулярной спектроскопии 3. Базовая и прикладная онтологии, Электронные библиотеки. 2012. т.15. в.2. URL: http://elbib.ru/index.phtml?page=elbib/rus/journal/2012/part2

[2] Tennyson J., Bernath P.F., Brown L.R., et al., IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part I. Energy Levels and Transition Wavenumbers for $H_2{}^{17}O$ and $H_2{}^{18}O$, J. Quant. Spectrosc. Radiat. Transfer, 2009, V. 110, Issue 9, P. 573-596.

[3] Tennyson J., Bernath P.F., Brown L.R., at al., IUPAC critical evaluation of the rotational–vibrational spectra of water vapor. Part II: Energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$, J. Quant. Spectrosc. Radiat. Transfer, 2010, V. 111, Issue 15, P. 2160-2184.

[4] Tennyson J., Bernath P.F., Brown L.R., et al., IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part III. Energy Levels and Transition Wavenumbers for $H_2{}^{16}O$, J. of Quant. Spectrosc. & Rad.Transfer, 2012. (in press)

[5] Polovtseva E. R., Voronina S.S., Lavrentiev N.A., et al., Information System for Molecular Spectroscopy. 5. Ro-vibrational Transitions and Energy Levels of the Hydrogen Sulfide Molecule / Atmos. and Oceanic Optics, 2012, v. 25, No. 2, p. 157–165.

[6] Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З. Базы знаний для описания информационных ресурсов в молекулярной спектроскопии. 2. Модель данных в количественной спектроскопии, Электронные библиотеки, 2011, т. 14, в.2, URL: http://elbib.ru/index.phtml?page=elbib/rus/journal/2011/part2

[7] Jacquinet-Husson N., Scott N.A., Chedin A. et al., The GEISA spectroscopic database: current and future archive for earth and planetary atmosphere studies, J. Quant. Spectrosc. & Rad. Transfer. 2008. v. 109. No 6. p. 1043-1059.

[8] Информационная система W@DIS. URL: http://wadis.saga.iao.ru

[9] Быков А.Д., Науменко О.Б., Синица Л.Н., Родимова О.Б., Творогов С.Д., Тонков М.В., Фазлиев А.З., Филиппов Н.Н., Информационные аспекты молекулярной спектроскопии, Томск, Из-во ИОА СО РАН, 2008, 360 с.

[10] Апанович З.В., Винокуров П.С., Ахлёстин А.Ю., Привезенцев А.И., Фазлиев А.З., Визуализация парных отношений источников данных в количественной спектроскопии, Материалы 15 Всеросс. Конф. «Интернет и современное общество», С-Петербург, 10-12 октября 2012, 2012, с. 7-15.

[11] Virtual atomic and molecular data centre / Dubernet M.L., at al. // J. Quant. Spectros. Rad. Transfer. 2010. v. 111. p. 2151–2159.