2013 | Volume 16| Issue 4|

Базы знаний для описания информационных ресурсов в молекулярной спектроскопии. 5. Качество экспертных данных

Ахлёстин А.Ю., Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З.

Аннотация

Показано, что доверие к контенту информационных ресурсов можно оценивать с помощью критерия опубликования и разделять ресурсы на доверяемую и сомнительную части. Задача оценки доверия состоит их четырех подзадач: (1) построения мультимножеств физических величин, содержащихся в первичных источниках данных, (2) согласования значений физических величин, (3) формирования количественных ограничений для критерия опубликования в разных интервалах изменения физических величин и (4) декомпозиции экспертных данных. Кратко описаны критерии достоверности спектральных данных и ограничения необходимые для решения задач согласования данных. Представлено табличное представление результатов согласования. На примере вакуумных волновых чисел описаны ограничения характерные для критерия опубликования. Оценки доверия, полученные из решения задачи декомпозиции, представлены в форме OWL-онтологий. Построение онтологической базы знаний подобного типа для виртуальных центров данных в дисциплинах с большими объемами данных измерений обеспечит автоматический выбор информационных ресурсов с высокой степенью доверия.

Ключевые слова: спектроскопия, согласование данных, доверие к контенту ресурсов, критерий опубликования.

Введение

Параметры спектральных линий применяются в разных предметных областях: оптике атмосферы, атмосферной радиации, астрономии и т.д. Потребность в таких данных постоянно растет, и увеличивается число производителей экспертных данных [1-11]. Более жесткими становятся требования к качеству данных: их точности, полноте и согласованности, достоверности и доверию. Показано, что для ряда прикладных задач существующие экспертные данные не адекватны требованиям, предъявляемым к ним, т.к. содержат в себе устаревшие, сомнительные и не полные наборы данных [12].

Экспертные информационные ресурсы количественной спектроскопии должны обеспечивать низкий порог адекватной передачи знаний о спектрах в прикладные науки. Последнее является важным в силу того, что исследователи прикладных наук не обладают знаниями необходимой глубины для понимания всех сторон

спектроскопических данных, и их выбор строится чаще на доверии, а не на собственной проверке достоверности данных. По этой причине исследователи должны иметь информацию о том, каким критериям достоверности и доверия удовлетворяют экспертные данные.

В нашей работе анализ качества экспертных спектральных данных сосредоточен на проверке их достоверности и оценке доверия по критерию опубликования [12]. Такой анализ является частью решения задачи оценки доверия экспертных информационных ресурсов по критерию опубликования. Эта задача разбивается на четыре подзадачи: (1) задачу построения мультимножеств значений физических величин, содержащихся в первичных источниках данных, (2) задачу согласования значений физических величин, (3) задачу формирования количественных ограничений для критерия опубликования в разных интервалах изменения физических величин и (4) задачу декомпозиции экспертных данных. Решение этих подзадач представлено в ИС W@DIS [13] в двух представлениях: табличном для исследователей и онтологическом для программных агентов.

Данная работа состоит из двух частей. В первой части работы рассмотрены критерии, по которым осуществляется согласование, и описаны интерфейсы пользователя для просмотра решений задачи согласования значений физических величин в спектроскопии. Мультимножества значений физических величин можно разделить на наборы канонических частей источников данных [14]. Для строгого решения задачи оценки доверия к экспертным ресурсам необходим полный набор опубликованных и достоверных значений физических величин и их согласование. В настоящее время достигнуты полнота и согласование спектральных данных для изотопологов молекулы воды [15-17], сероводорода [18] и монооксида углерода [19]. В ИС W@DIS размещены спектральные характеристики молекул аммония [20], диоксида углерода [21], закиси азота, карбонил сульфида [22] и фосфина [23], опубликованные более чем в 20 журналах за период 80 лет. С определенной осторожностью можно утверждать о полноте собранных нами данных. Онтология информационных ресурсов по количественной спектроскопии [24] описывает состояние дел с достоверностью и согласованностью ресурсов по перечисленным выше молекулам.

Во второй части работы сформулированы задачи об ограничениях в критерии опубликования и декомпозиции, описан табличный интерфейс для просмотра решения задачи декомпозиции, а также T-box и A-box онтологической базы знаний, содержащие оценки доверия к экспертным ресурсам о спектральных характеристиках ряда молекул. Структура программного обеспечения для решения задачи декомпозиции описана в [25]. В работе [12] был предложен критерий опубликования для оценки доверия к экспертным данным в количественной спектроскопии. В этой работе количественные ограничения ai для критерия опубликования в шести интервалах, относящихся к диапазону 0-30000 $см^{-1}$, являлись параметрами подгонки. Они подбирались таким образом, чтобы процент сомнительных переходов в экспертных данных не превышал 10% при условии, что изменение величины a_i проводилось кратно десяти. В данной работе выбор величин a_iопределялся близостью к разрешающей способности измерительной аппаратуры для 12 интервалов от радио-частот до рентгеновского излучения. В этой части работы описана структура индивида, характеризующего

результаты оценки доверия, дан пример формирования класса **DecompositionNearInfraredRangeDescription**, и приведена статистика сомнительных волновых чисел для шести изотопологов молекулы диоксида углерода.

При создании онтологий использовался традиционный методологический подход, согласно которому существующие модели данных и метаданных исследуемых предметных областей ассимилируются в создаваемые онтологии. При этом, эти модели анализируются и расширяются (или редуцируются) до необходимого уровня грануляции соответствующей предметной области.

1. Предметные задачи молекулярной спектроскопии

Большую часть информационных ресурсов в спектроскопии представляют решения шести задач [26]. На рис.1 схематически показаны такие решения, объединенные в группы, и процедуры с помощью которых они получаются. Прежде всего, выделены группы решений, связанные с измерениями и вычислениями. Результатом измерений являются мультимножества физических величин, характеризующих состояния и переходы молекул. Прямые и обратные связи на рис.1 показывают, что процесс получения всех четырех типов массивов данных является итерационным. Связано это с тем обстоятельством, что квантовые числа, однозначно описывающие состояния и переходы, не являются измеримыми величинами. Они появляются в результате расчетов, для проведения которых используются параметры, определяемые из измерений. Правило Ридберга-Ритца позволяет находить эталонные переходы, используя для этого уровни энергии, вычисляемые из мультимножества вакуумных волновых чисел.

Мультимножества измеренных уровней энергии, волновых чисел и параметров спектральных линий применяются для вычисления характеристик эталонных и предсказанных переходов и состояний, а также экспертных данных. Значения физических величин, входящие в мультимножества, составляют основную часть в публикациях, относящихся к молекулярной спектроскопии. Согласованность этих значений, их достоверность и доверие к ним лежат в основании качества всех информационных ресурсов молекулярной спектроскопии.

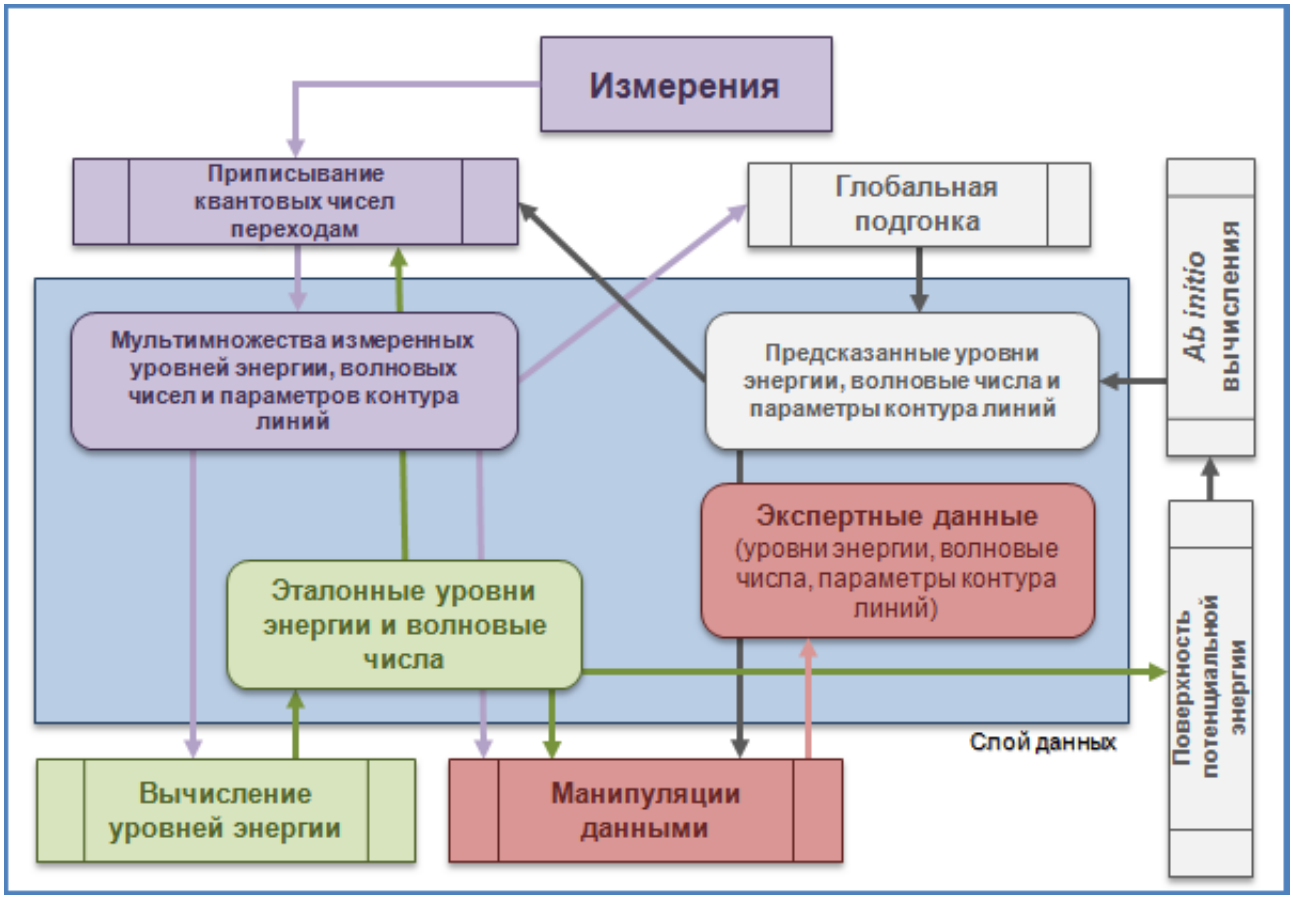


Рис. 1. Схематичное представление отношений между приложениями и связанными с ними входными и выходными данными.

Среди четырех видов данных необходимо выделить экспертные данные. Для выделения есть две причины. Во-первых, экспертные данные подготавливаются для прикладных предметных областей, число которых превышает несколько десятков, и каждая из которых имеет свои требования к качеству данных. Во-вторых, экспертные данные получаются в результате неформальных манипуляций данными, при которых квалификация и пристрастие экспертов играют существенную роль. По этой причине, при исследовании качества экспертных данных необходимо использовать критерии доверия в их различной форме. Ниже для этой цели используется критерий опубликования [12].

2. Качество контента информационных ресурсов количественной спектроскопии

В рамках подхода Semantic Web (SW) контент информационных ресурсов включает в себя данные, информацию и знания. Эти термины в литературе трактуются достаточно широко. Мы будем следовать складывающейся терминологии SW, используя вместо них термины «данные», «связанные данные» и «онтология». В SW интерпретация терминов «данные», «связанные данные» и «онтология» тесно связана с семантикой формальных языков XML, RDF и OWL. Более того, такой интерпретацией обусловлено использование терминов «слой данных и приложений», «информационный слой» и «слой знаний», введенных в eScience [27] для описания инфраструктуры информационных ресурсов. В частности, ИС

W@DIS, описанию ресурсов которой посвящена наша работа, имеет подобную трехслойную архитектуру [25, 28].

В этой работе рассматривается качество экстенсионала данных. Основной акцент сделан на исследовании доверия и достоверности экстенсионала данных. Здесь под достоверностью понимается соответствие данных формальным ограничениям (критериям), следующим из математических моделей молекул и условий согласования идентичных канонических частей источников спектральных данных. Критерий опубликования применяется к каноническим частям экспертных источников данных и позволяет выделить в них сомнительные данные, которые могут быть неопубликованными данными или данными не согласованными с первичными данными. Исследователи прикладных предметных областей, использующие экспертные спектральные данные, должны проверять сомнительные данные по критериям соответствующих предметных областей.

2.1. Достоверность спектральных данных

Для решения задач глобальной подгонки, вычисления эталонных уровней энергии и построения экспертных данных необходимо выделить достоверные данные в мультимножестве измеренных данных. Рассмотрим две группы критериев, определяющих достоверность спектральных данных. Первая группа включает в себя критерии, следующие из выбора математических и физических моделей молекул. Вторая группа критериев связана с ограничениями, обусловленными процедурами согласования результатов измерений и вычислений характеристик идентичных состояний и переходов.

Экспертные и первичные источники спектральных данных содержат характеристики переходов и состояний молекул. Они могут содержать переходы и состояния, характеристики которых не удовлетворяют формальным ограничениям, следующим из моделей молекул. По этой причине их разделяют на две части: каноническую и не каноническую. Каноническая часть содержит переходы (состояния), обладающие полным набором квантовых чисел, удовлетворяющем правилам отбора (ограничениям на квантовые числа состояния), и не содержит переходов (состояний) с одинаковым набором квантовых чисел. Ниже, рассматривая экспертные и первичные данные, будем иметь в виду только их каноническую часть. В ИС W@DIS мультимножество измеренных переходов (состояний) формируется из канонических частей первичных источников данных, а мультимножество предсказанных переходов (состояний) может содержать неканоническую часть.

Канонические части измеренных данных необходимо согласовывать между собой. В спектроскопии объективными причинами рассогласования данных являются различия в точности измерений, обусловленными физическими причинами, в частности, калибровкой аппаратуры, и неоднозначность процедуры приписывания квантовых чисел. Такого рода рассогласования можно избежать перекалибровкой данных и переприписыванием квантовых чисел. Оставшиеся рассогласованные данные, как правило, исключаются экспертами из канонической части. Количественные ограничения, используемые при согласовании, различны для разных областей изменения значений физических величин.

2.2. Доверие к экспертным информационным ресурсам

В литературе много интерпретаций понятия «доверие» [29-34], и, соответственно, много критериев по которым оно оценивается. Общим для большинства интерпретаций является то, что критерии доверия представляют собой частично формализованные ограничения. Предложенный в [12] критерий опубликования относится к группе критериев оценки доверия к контенту информационных ресурсов [33, 34], и позволяет выделять сомнительную часть контента. Применение критерия опубликования сводится к проверке совпадения значений идентичных параметров спектральных линий экспертного массива данных с опубликованными первичными данными в рамках заданной точности.

С первых дней развития Semantic Web (SW) понятие «доверие» является центральным в этом подходе. Оно указывает его цель – получения пользователем сети Интернет достоверных информационных ресурсов, обладающих высокой степенью доверия. Исследователь, работая с ресурсами, имеет возможность проверять достоверность ресурсов в рамках своей компетенции. Более того его профессиональные навыки и опыт позволяют ему принимать решения в тех случаях, когда у него для оценок имеются частично формализованные критерии для проверок, а оценки требуют ресурсы, обладающие неопределенностью. Подход SW ориентирован не только на исследователей, но и на агентов которые должны решать такие задачи. «В SW, где контент представлен в онтологиях и аксиомах, как компьютеру решить какому из источников доверять, если все они противоречат друг другу.» [33].

Наряду с оценкой доверия к ресурсам на основе знаний об их производителях, о создателях агентов и технологий их создания в последние несколько лет появились работы [33,34] о доверии к контексту ресурсов. В них выделены 19 факторов, влияющих на доверие к контенту. Эти факторы можно разбить на две группы. В первую группу входят факторы, не зависящие от социальных факторов или личности исследователя, такие, как соответствие информационного ресурса его назначению, качество контента и его частей, наличие альтернативных ресурсов, непредвзятость, ограниченность ресурсов, наличие подобных ресурсов, детализация, возраст, корректность представления, новизна и отсутствие жульничества. Ко второй группе относятся популярность, авторитетность и рекомендуемость ресурсов, собственный опыт и способность проведения экспертизы, побуждения и способность пользователя к компромиссу.

Экспертные данные в количественной спектроскопии удовлетворяют ряду факторов доверия. Они популярны, авторитетны, рекомендуемы специалистами в прикладных областях, они производятся несколькими группами экспертов, обладают корректным представлением и не содержат намеренно размещенной некорректной информации. Однако существует факторы, заставляющие сомневаться в корректности части данных. Доля сомнительных данных в значительной степени зависит от декларации назначения конкретных экспертных данных. Как правило, эксперты не определяют строго области применимости спектральных данных. В наиболее популярных ресурсах [2, 7, 8] незначительные части ресурсов не достоверны, происхождение части данных неизвестно (как правило, они не опубликованы), возраст некоторых сомнительных данных

значителен, и механизмы обратной связи пользователей с экспертами практически отсутствуют. Перечисленные недостатки указывают на актуальность оценок доверия экспертным данным в количественной спектроскопии.

3. Достоверность ресурсов по количественной спектроскопии

3.1. Критерии достоверности источников данных

В спектроскопии к числу ограничений, определяющих каноническую часть источников данных, следующих из математической модели молекул, относятся ограничения на квантовые числа состояний и переходов, так называемые ограничения на квантовые числа состояний и правила отбора. Эти ограничения проверяются в ИС W@DIS при импорте данных и формировании источника данных. Другими ограничениями, определяющими каноническую часть источника данных, являются требования отсутствия дубликатов состояний и переходов и отсутствие переходов и состояний, характеризуемых неполным набором квантовых чисел.

3.1. Критерии достоверности источников данных

3.1.1. Правила отбора и ограничения на состояния

Правила отбора и ограничения на квантовые числа состояний представляют собой линейные зависимости между квантовыми числами описываемые равенствами и неравенствами. Примеры правил отбора и их систематизация приведены в [24] (см. Таблица 9. Реестр правил отбора …).

3.2. Критерии согласования данных в парах первичных источников данных

Переходы и состояния в спектроскопии описываются наборами значений физических величин. Квантовые числа однозначно характеризуют переходы и состояния молекулы, но не являются измеримыми величинами. Остальные физические величины могут быть извлечены из данных измерений. Если квантовые числа переходов одинаковы, то такие переходы называют идентичными. При сравнении источников данных рассматриваются пары идентичных переходов или состояний из их канонических частей. Для согласования данных необходимо достичь выполнения неравенств на максимальную разность физических величин и среднеквадратических отклонений, а также нулевых значений показателей схожести.

3.2.1. Максимальная разность значений физических величин при сравнении пары источников данных

Значения измеренных физических величин идентичных переходов могут различаться. Если для каждой пары идентичных переходов определить разности физических величин F_i, а затем найти максимальные разности ΔF_{i}, то ΔF_{i} могут использоваться как индикатор необходимости согласования. Приведем пример.

Современные измерения положений центров линий в инфракрасном диапазоне проводятся с точностью до 0.001 см-1. Если положить, что критическим значением разности волновых чисел является 0.035 $см^{-1}$, то идентичные переходы с $\Delta F_i >$ 0.035 не согласованы, и причиной может быть некорректное приписывание квантовых чисел. На практике используют также другое критическое значение ΔF_i при превышении которого можно утверждать, что рассогласование не достижимо путем изменения квантовых чисел. Следовательно, величина ΔF_i позволяет выделять наиболее грубые ошибки в приписывании квантовых чисел.

3.2.2. Среднеквадратическое отклонение

В спектроскопии используется среднеквадратическое отклонение для сравнения двух источников данных, ни один из которых нельзя считать «эталоном». Списки переходов в источниках данных можно представить в виде векторов

$A_1=(x^1_{qn1}, x^1_{qn2}, \ldots, x^1_{qnm})$ и $A_2=(x^2_{qn1}, x^2_{qn2}, \ldots, x^2_{qnm})$,

где x^i_{qnj} – значения физических величин (например, вакуумных волновых чисел), а qn_j – соответствующие им квантовые числа, m - число идентичных переходов. Формула для вычисления среднеквадратического отклонения имеет вид:

$$RMSD(A_1, A_2)=(\sum^{qnm}_{i=qn1}(x^1_i- x^2_i)^2/m)^{1/2}. \quad (1)$$

В некоторых представлениях квантовые числа можно разбить на две группы. Одна группа называется колебательными квантовыми числами, а другая – колебательно-вращательными квантовыми числами. Ниже такое разделение представлено в виде

$qn=(qn_{vib}, qn_{v\text{-}r})$.

При описании переходов вводят понятие «колебательная полоса спектральных линий» которому соответствует набор спектральных линий с разными наборами колебательно-вращательных квантовых чисел, но с фиксированными значениями колебательных квантовых чисел . В количественной спектроскопии часто используют формулу для расчета, в которой среднеквадратическое значение рассчитывается для каждой колебательной полосы. В этом случае в формуле (1) суммирование ведется только по колебательно-вращательным квантовым числам выбранной полосы. Для описания такого отклонения используется термин «среднеквадратическое отклонение по колебательной полосе», которое определяется формулой

$$RMSD(A_1(qn_{vib}, qn_{v\text{-}r}), A_2(qn_{vib}, qn_{v\text{-}r}))\ (qn_{vib}=const) = (\sum^k_{i=1}(x^1_i- x^2_i)^2/k)^{1/2}. \quad (2)$$

В формуле (2) k представляет число идентичных вращательных переходов полосы. Введенное выше среднеквадратическое отклонение, заданное формулой (1), будем называть отклонением по всем полосам или интегральным среднеквадратическим отклонением.

3.2.3. Сравнение упорядочений

При анализе решений задач спектроскопии можно добиться требуемой малости значений разности физических величин идентичных переходов и требуемых значений среднеквадратических отклонений. Однако этого бывает недостаточно для согласования решений задач спектроскопии.

Пусть есть два набора A_1(a1i,qi) и A_2 (a2i,qi) значений физических величин (вакуумных волновых чисел, интенсивностей, и т.д.) и соответствующих им квантовых чисел. При этом положим, что набору квантовых чисел в A1 соответствует тождественный ему набор квантовых чисел в A2. Упорядочим последовательности значений физической величины по возрастанию. Припишем порядок возрастания, относящийся к физической величине, квантовым числам. В общем случае порядок следования квантовых чисел в двух наборах может быть различным. Например, сравнение двух экспертных массивов [2,8] показывает, что он нарушается для более 700 переходов в случае молекулы воды. Заметим, что эти массивы имеют более тридцати тысяч идентичных переходов.

Количественная оценка разупорядочения квантовых чисел в паре источников данных может определяться разными способами. В нашей работе определены и вычисляются три таких показателя. Показателем схожести A00 называется минимальное число перестановок переходов в одном из источников данных, необходимое для совпадения порядка следования квантовых чисел в нем с порядком следования квантовых чисел в другом источнике данных. Показателем схожести A01 первого источника данных со вторым источником данных называется число исключённых переходов из обоих источников данных в соответствии с алгоритмом, описанным в Приложении.

На Рис.2 представлено графическое сравнение упорядочений квантовых чисел. Оно состоит из трёх цветных полос. Средняя полоса представляет собой правильный порядок следования квантовых чисел в одном из двух сравниваемых источников данных. Верхняя и нижняя полосы соответствуют источникам данных, библиографические ссылки на которые размещены сверху и снизу полос, соответственно.

Предположим, что порядок следования квантовых чисел правильный в работе 1985_CaFlMaCh. Сравниваемые источники данных имеют 18 идентичных переходов. Каждому переходу, идентифицируемому квантовыми числами, соответствует определенный цвет. Цветовое представление верхней полосы формируется следующим образом. Упорядочим переходы в источнике данных 2006_NaSnTaSh по значениям волновых чисел. Припишем переходам этого источника данных цвета, которые соответствуют цветам источника данных 1985_CaFlMaCh. Значение величины A10 указывает, какое минимальное число переходов надо удалить в верхней и средней полосе, чтобы порядок цветов соответствовал естественному порядку.

Рис.2. Визуализация сравнения упорядоченных по значению вакуумных волновых чисел.

На рис.2 значение A_{10} равно 5. Черные прямоугольники между верхней и нижней полосой указывают на конкретные переходы, которые следует удалить.

Аналогичным образом, если считать правильным порядок следования квантовых чисел в 2006_NaSnTaSh, то можно посчитать величину A01. Нижняя полоса представляет упорядоченный по значению вакуумных волновых чисел источник данных 1985_CaFlMaCh, в котором переходам приписан цвет, соответствующий порядку их следования в источнике данных 2006_NaSnTaSh.

3.3. Табличное представление результатов согласования

При импорте данных в ИС W@DIS вычисляются все показатели, характеризующие согласование данных. Эти показатели связаны со свойствами, описанными в онтологии информационных ресурсов [24]. В настоящее время нет программного обеспечения, позволяющего проводить вычисления в рамках OWL-онтологий. Согласование данных исследователями, применяющими спектральные данные в прикладных предметных областях, основывается на разных количественных ограничениях на величины максимальной разности, среднеквадратичного

отклонения и показателей схожести. В информационной системе учет особенностей прикладных областей приводит к созданию онтологий пользователей, о которых пойдет речь в шестой части этой серии статей. Здесь мы приведем описание интерфейсов для просмотра результатов согласования данных, в которых эти ограничения можно задавать пользователям, и в табличном виде получать результаты согласования при заданных значениях ограничений.

Подобное табличное представление показано на рис. 3 и 4. Оно удобно при детальном анализе качества данных и на этапе выбраковки некачественных данных, но требует значительного времени для анализа. Более того, табличное представление среднеквадратических отклонений по отдельным полосам является громоздким и необозримым, и оно не реализовано в ИС W@DIS.

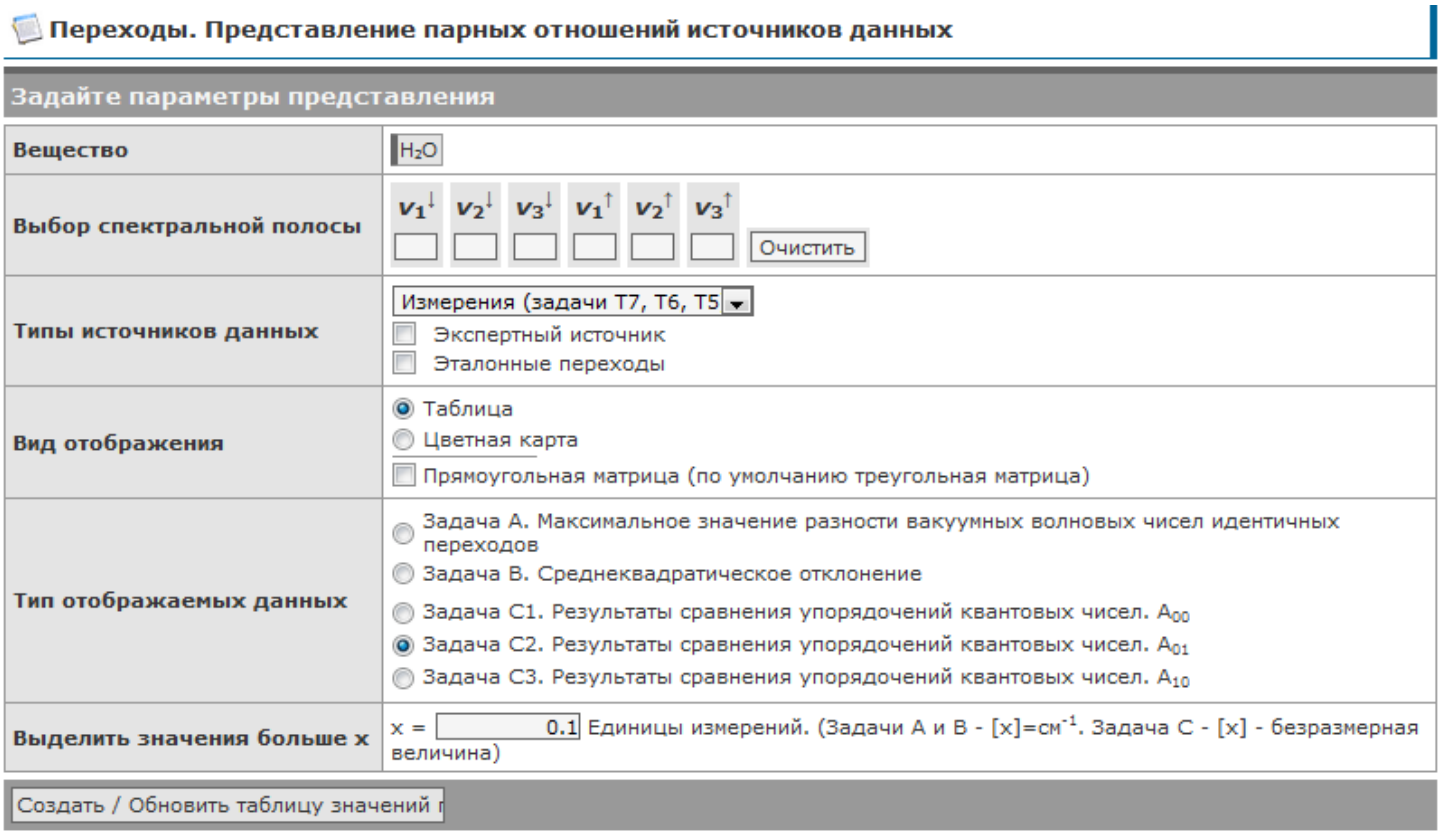


Рис. 3.А). Интерфейс для задания типа бинарного отношения, определяемого задачей (A, B, C), и параметров, характеризующие представление данных.

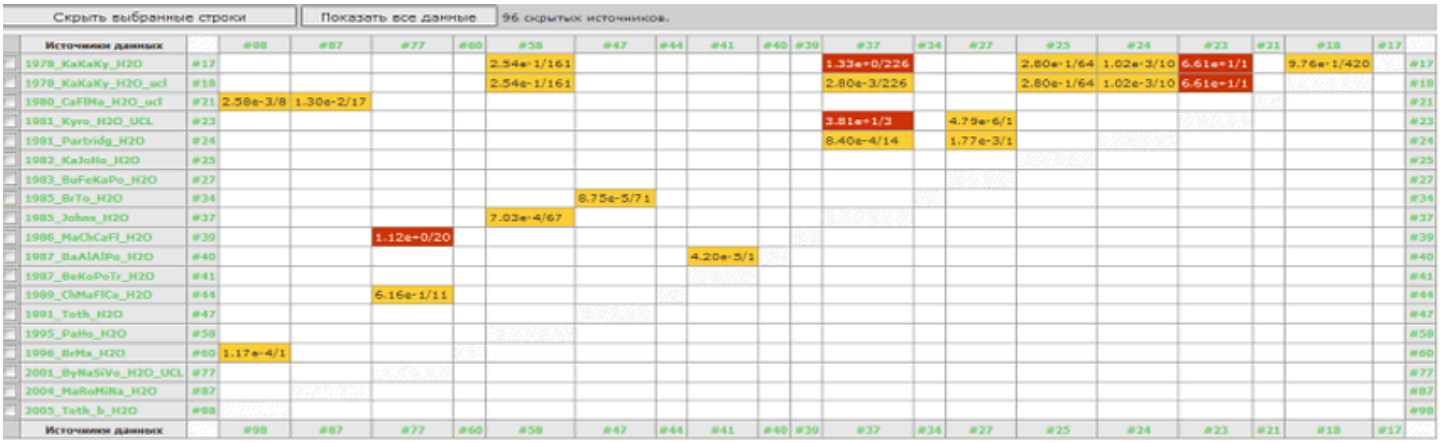
На рис.3А показан интерфейс для выбора молекулы, источников данных определенного типа, способа представления и результатов согласования по определенной задаче. Параметром x можно задать критическое значение, при превышении которого выделяются пары источников данных, не удовлетворяющие такому ограничению. Представление результатов для трех задач показаны ниже на рис.3Б-Г. Если исследователю необходим просмотр результатов согласования данных для отдельной спектральной полосы, он может достичь этого выбором соответствующей полосы в представленном интерфейсе.

Из анализа результатов согласования, представленных на этих рисунках, следует, что данные могут быть согласованы по одним показателям, но не согласованы по

другим. Например, источники данных 1980_CaFlMa и 2004_MaRoMiNa согласованы по максимальной разности волновых чисел и интегральному среднеквадратическому отклонению, но имеют слабое разупорядочение квантовых чисел. Источники 1978_KaKaKy и 1995_PaHo согласованы по показателю схожести и среднеквадратическому отклонению, но не согласованы по максимальной разности. Источники 1986_MaChCaFl и 2001_ByNaSiVo не согласованы по максимальной разности и среднеквадратическому отклонению, но согласованы по показателю схожести.

Источники данных		#98	#87	#77	#60	#58	#47	#44	#41	#40	#39	#37	#34	#27	#25	#24	#23	#21	#18	#17	
1978_KaKaKy_H2O	#17					3.00e+0/161						2.00e+1/226			2.00e+0/64	2.00e-3/10	6.61e+1/1		2.00e+1/420		#17
1978_KaKaKy_H2O_ucl	#18					3.00e+0/161						2.41e-2/226			2.00e+0/64	2.00e-3/10	6.61e+1/1				#18
1980_CaFlMa_H2O_ucl	#21	4.76e-3/8	3.60e-2/17																		#21
1981_Kyro_H2O_UCL	#23											6.61e+1/3		4.79e-6/1							#23
1981_Partridg_H2O	#24											1.84e-3/14		1.77e-3/1							#24
1982_KaJoHo_H2O	#25																				#25
1983_BuFeKaPo_H2O	#27																				#27
1985_BrTo_H2O	#34						5.13e-4/71														#34
1985_Johns_H2O	#37					4.44e-3/67															#37
1986_MaChCaFl_H2O	#39			3.54e+0/20																	#39
1987_BaAlAlPo_H2O	#40								4.20e-5/1												#40
1987_BeKoPoTr_H2O	#41																				#41
1989_ChMaFlCa_H2O	#44			2.04e+0/11																	#44
1991_Toth_H2O	#47																				#47
1995_PaHo_H2O	#58																				#58
1996_BrMa_H2O	#60	1.17e-4/1																			#60
2001_ByNaSiVo_H2O_UCL	#77																				#77
2004_MaRoMiNa_H2O	#87																				#87
2005_Toth_b_H2O	#98																				#98
Источники данных		#98	#87	#77	#60	#58	#47	#44	#41	#40	#39	#37	#34	#27	#25	#24	#23	#21	#18	#17	

3.Б) Максимальное значение разности вакуумных волновых чисел идентичных переходов



Скрыть выбранные строки | Показать все данные | 96 скрытых источников.

Источники данных		#98	#87	#77	#60	#58	#47	#44	#41	#40	#39	#37	#34	#27	#25	#24	#23	#21	#18	#17	
1978_KaKaKy_H2O	#17					2.54e-1/161						1.33e+0/226			2.80e-1/64	1.02e-3/10	6.61e+1/1		9.76e-1/420		#17
1978_KaKaKy_H2O_ucl	#18					2.54e-1/161						2.80e-3/226			2.80e-1/64	1.02e-3/10	6.61e+1/1				#18
1980_CaFlMa_H2O_ucl	#21	2.58e-3/8	1.30e-2/17																		#21
1981_Kyro_H2O_UCL	#23											3.81e+1/3		4.79e-6/1							#23
1981_Partridg_H2O	#24											8.40e-4/14		1.77e-3/1							#24
1982_KaJoHo_H2O	#25																				#25
1983_BuFeKaPo_H2O	#27																				#27
1985_BrTo_H2O	#34						8.75e-5/71														#34
1985_Johns_H2O	#37					7.03e-4/67															#37
1986_MaChCaFl_H2O	#39			1.12e+0/20																	#39
1987_BaAlAlPo_H2O	#40								4.20e-5/1												#40
1987_BeKoPoTr_H2O	#41																				#41
1989_ChMaFlCa_H2O	#44			6.16e-1/11																	#44
1991_Toth_H2O	#47																				#47
1995_PaHo_H2O	#58																				#58
1996_BrMa_H2O	#60	1.17e-4/1																			#60
2001_ByNaSiVo_H2O_UCL	#77																				#77
2004_MaRoMiNa_H2O	#87																				#87
2005_Toth_b_H2O	#98																				#98
Источники данных		#98	#87	#77	#60	#58	#47	#44	#41	#40	#39	#37	#34	#27	#25	#24	#23	#21	#18	#17	

3.В) Интегральное среднеквадратическое отклонение

Источники данных		#98	#87	#77	#60	#58	#47	#44	#41	#40	#39	#37	#34	#27	#25	#24	#23	#21	#18	#17	
1978_KaKaKy_H2O	#17					0/161						1/226			0/64	0/10	0/1		1/420		#17
1978_KaKaKy_H2O_ucl	#18					0/161						0/226			0/64	0/10	0/1				#18
1980_CaFlMa_H2O_ucl	#21	0/8	1/17																		#21
1981_Kyro_H2O_UCL	#23											1/3		0/1							#23
1981_Partridg_H2O	#24											0/14		0/1							#24
1982_KaJoHo_H2O	#25																				#25
1983_BuFeKaPo_H2O	#27																				#27
1985_BrTo_H2O	#34						0/71														#34
1985_Johns_H2O	#37					0/67															#37
1986_MaChCaFl_H2O	#39			0/20																	#39
1987_BaAlAlPo_H2O	#40								0/1												#40
1987_BeKoPoTr_H2O	#41																				#41
1989_ChMaFlCa_H2O	#44			0/11																	#44
1991_Toth_H2O	#47																				#47
1995_PaHo_H2O	#58																				#58
1996_BrMa_H2O	#60	0/1																			#60
2001_ByNaSiVo_H2O_UCL	#77																				#77
2004_MaRoMiNa_H2O	#87																				#87
2005_Toth_b_H2O	#98																				#98
Источники данных		#98	#87	#77	#60	#58	#47	#44	#41	#40	#39	#37	#34	#27	#25	#24	#23	#21	#18	#17	

3.Г) Результаты сравнения упорядочений (коэффициент A01)

Рис.3. Фрагмент таблицы, представляющий не согласованные пары первичных источников измеренных данных, относящихся к основному изотопологу молекулы воды.

На Рис.3 приводятся результаты согласования 21 источника данных по величине максимальной разности волновых чисел, среднеквадратичному отклонению и показателю схожести. В этом фрагменте таблицы только 12 источников данных содержат переходы, для которых можно найти идентичные, хотя бы в одном из оставшихся 11 источников. В таблице приведены аббревиатуры публикаций, располагающиеся в таблице в хронологическом порядке, и в соответствующей ячейке таблицы приведено отношение двух чисел. Числитель отношения указывает значение бинарного отношения, а знаменатель – число идентичных переходов в этих источниках данных. Источники данных, не содержащие идентичных переходов с другими публикациями, в таблице не приведены. Жёлтый цвет ячейки таблицы означает, что соответствие между идентичными переходами удовлетворяет заданному критерию, а красный цвет – не удовлетворяет критерию.

Для решения некоторых задач анализа качества данных можно ограничиться качественным просмотром результатов согласования данных. В этом случае из таблиц удаляются количественные характеристики, и вводят цветовое разделение на пары, удовлетворяющие количественным критериям и неудовлетворяющие им. Представленная на Рис.4 таблица содержит значительное число несогласованных вакуумных волновых чисел в источниках данных и является симметричной относительно диагонали.

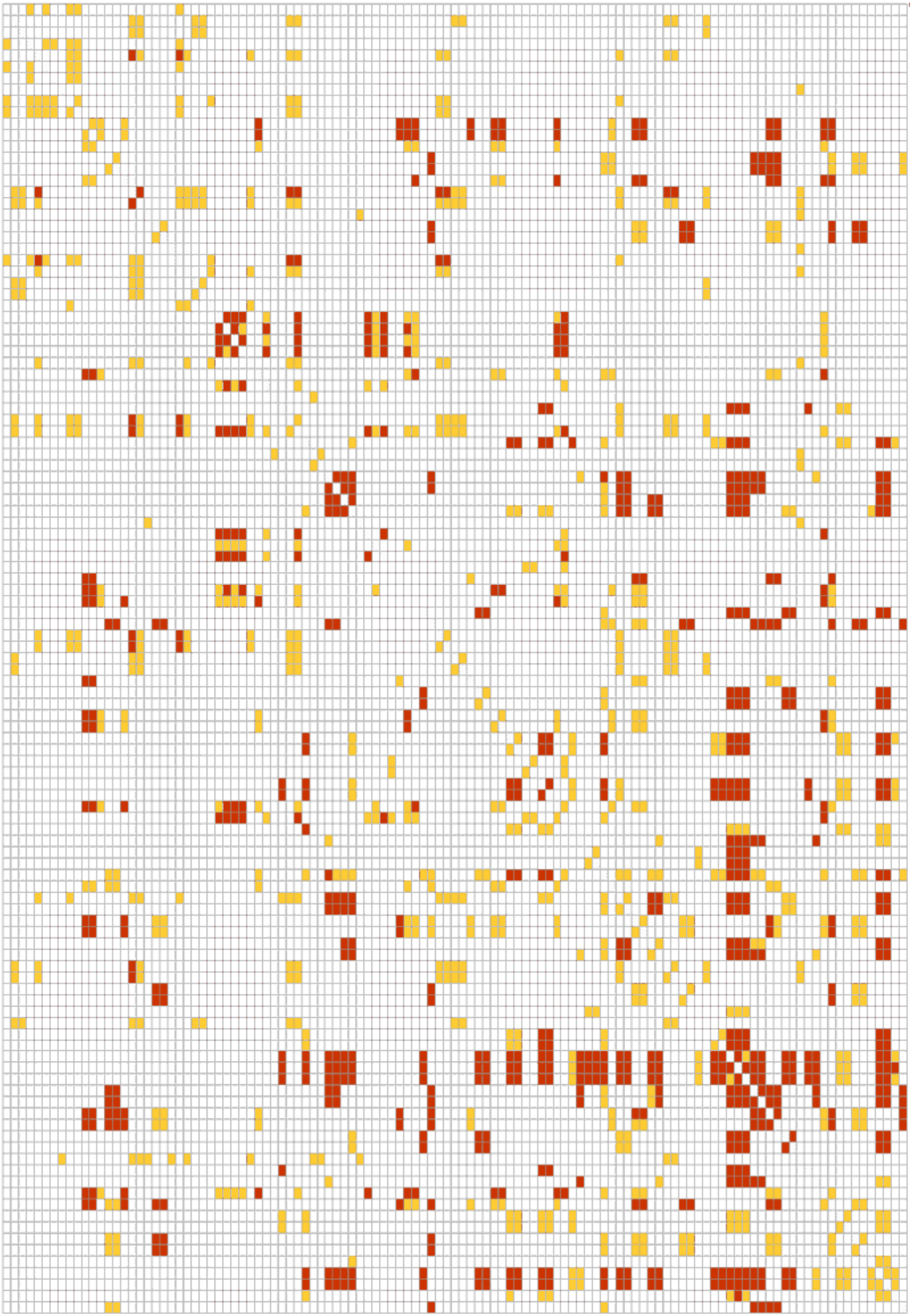
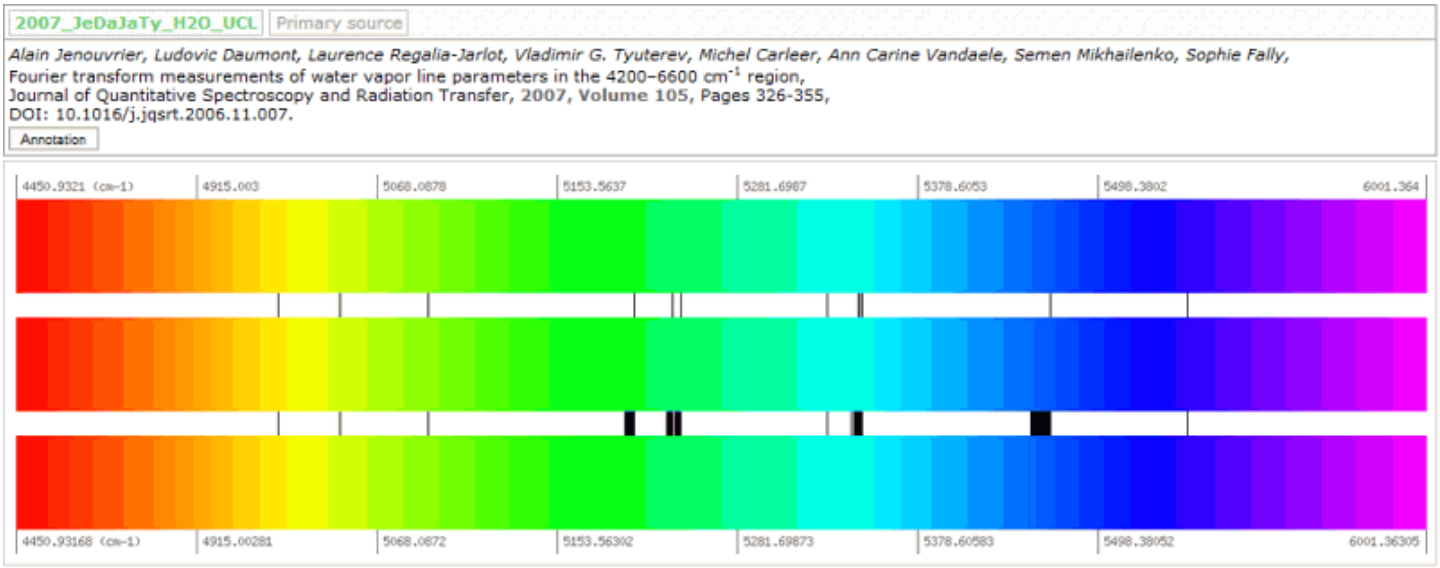


Рис.4. Цветная карта для анализа среднеквадратических отклонений для первичных экспериментальных источников информации по молекуле CO2. (Обозначения: - интегральное среднеквадратическое отклонение меньше 0.01 см-1, - среднеквадратическое отклонение больше 0.01 см-1).

С каждой цветной ячейкой таблицы связан рисунок, обеспечивающий исследователя более детальной информацией о значениях величин, определяющих согласование источников данных.



Comparison of the Data Sources

The Number of Transitions in Identical Ro-Vibrational Bands	$\Delta\omega_{max}$	Vacuum Wavenumbers RMSD	A_{00}	A_{01}	A_{10}
1212	1.12e+1	7.62e-1	82	54	14

Comparison of the Data Sources on Separated Ro-Vibrational Bands

$v_1^{\downarrow}$	$v_2^{\downarrow}$	$v_3^{\downarrow}$	$v_1^{\uparrow}$	$v_2^{\uparrow}$	$v_3^{\uparrow}$	The Number of Transitions in Identical Ro-Vibrational Bands	$\Delta\omega_{max}$	Vacuum Wavenumbers RMSD	A_{00}	A_{01}	A_{10}
0	1	1	0	0	0	471	1.12e+1	1.12e+0	58	24	9
1	1	0	0	0	0	381	7.83e+0	5.36e-1	21	16	2
0	2	1	0	1	0	173	1.05e-2	1.19e-3	1	8	2
0	3	0	0	0	0	154	3.70e-3	5.90e-4	0	4	0
1	2	0	0	1	0	33	1.50e-3	5.90e-4	2	2	1

Рис.5. Детальная информация о величинах, характеризующих пару источников данных (2007_JeDaJaTy, 2005_Toth).

На рисунке явно выделено в каких участках спектра наблюдается нарушения порядка следования переходов, и для каждой пары идентичных полос указаны значения ??max, RMSD и показатели схожести. Рассогласование данных по максимальной разности и RMSD присуще только двум полосам из пяти.

Наконец отметим две наши работы [35, 36] в которых представлены в графическом виде результаты рассогласования в виде графов цвет дуг которых определяет согласованность или не согласованность пар источников данных.

4. Оценка доверия к экспертным данным по критерию опубликования

Задача оценки доверия к экспертным данным состоит в следующем: Пусть физическая величина PQ характеризует состояние или переход в физической системе, а уникальность состояний или переходов определяется фактором qn_i. Пусть имеется множество выбранных экспертами значений физической величины $A = \{PQ(p^i, qn_i)\}$, в котором все qni уникальны. Пусть $M(p^i_j, qn_j)$ - мультимножество всех опубликованных измеренных значений PQ, содержащее только достоверные значения, а $T(p^i_j, qn_j)$ – мультимножество опубликованных расчетных значений PQ. Пусть на интервале изменения значений PQ из A погрешности измерения Di, разные в силу специфики принципов функционирования приборов (или существуют иные причины дифференциации по величине значений PQ для определения допустимой критерием опубликования величины отклонения экспертных значений PQ от идентичные значений PQ из $M(p^i_j, qn_j)$ и $T(p^i_j, qn_j)$). Найти число значений PQ из A, удовлетворяющих (E) и не удовлетворяющих (F) критерию опубликования и области изменения их значений в каждом из интервалов изменения PQ, характеризуемым Di при условии, что процедура выбора значений экспертами не является явной и контролируемой.

В этом параграфе рассмотрены две задачи: задача формирования количественных ограничений для критерия опубликовании в разных интервалах изменения физических величин и задача декомпозиции экспертных данных. На примере молекулы диоксида углерода описано табличное и онтологическое представление результатов решения задачи декомпозиции.

4.1. Формирование количественных ограничений для критерия опубликования в количественной спектроскопии

Различают два вида критериев опубликования: строгий и слабый. Строгий критерий опубликования требует, при сравнении идентичных переходов экспертного и первичного массива данных, равенства значений физических величин. Слабый критерий основан на проверке неравенства $| p_{expert} – p_{primary}| < p_0$, где p_{expert} , $p_{primary}$ - значения идентичных физических величин (экспертной и первичной), а p_0 – допустимая критерием величина отклонения. В спектроскопии используется слабый критерий опубликования. Более того, число p_0 зависит от области изменения значений физической величины, по которой осуществляется декомпозиция информационных ресурсов.

В этой работе рассмотрена декомпозиция только по вакуумным волновым числам для которых использованы отклонения значения которых в радио-частотном интервале 10^{-6}, в микроволновом – 10^{-5} , в дальнем ИК, в длинноволновом ИК и в среднем ИК – $5*10^{-3}$, в коротковолновом – $5*10^{-2}$, в ближнем ИК и видимом – $5*10^{-3}$.

4.2. Декомпозиции экспертных данных

Экспертные данные в количественной спектроскопии содержат физические величины, характеризующие переходы и состояния, и библиографические ссылки,

указывающие, из каких публикаций извлечены значения этих величин. Переходы в количественной спектроскопии описываются набором физических характеристик. К их числу относятся квантовые числа, вакуумные волновые числа, интенсивности, уровни энергии и параметры контура спектральной линии.

Для решения задачи декомпозиции используются входные данные, к числу которых относятся мультимножества опубликованных первичных экспериментальных и теоретических данных. Оценки доверия к экспертным данным включает в себя результаты декомпозиции по каждой физической величине или по всем физическим величинам, удовлетворяющим критериям опубликования для каждой из них. Ниже рассмотрены результат декомпозиции для одной физической величины – вакуумных волновых чисел.

Значимыми для декомпозиции экспертного массива данных являются полнота и согласованность мультимножеств физических величин и величины Dj. Полнота предполагает присутствие в мультимножестве всех опубликованных канонических первичных данных, но допускается случай, когда в нем не присутствуют первичные источники, являющиеся частями включенного в него первичного источника. Параметром, характеризующим количественную характеристику полноты, является список всех первичных источников, использованных при декомпозиции. Согласованность данных определяется количественными характеристиками, описывающими бинарные отношения между идентичными частями источников данных измерений, входящих в мультимножество. При декомпозиции используются не согласованные между собой вычисленные данные, объём которых составляет несколько миллионов записей для каждой из молекул. Для согласования первичных измеренных данных выбраны бинарные отношения между идентичными переходами и колебательно-вращательными полосами. Согласование идентичных переходов определяется по двум типам характеристик: максимально допустимой разнице между значениями идентичных переходов в экспертном массиве и показателям разупорядочения квантовых чисел в сравниваемых данных. Согласование между колебательно-вращательными полосами определяется допустимой величиной их интегрального среднеквадратического отклонения и среднеквадратического отклонения по отдельным колебательно-вращательным полосам. Учитывая тот факт, что большая часть первичных измеренных данных в количественной спектроскопии не согласована (исключение составляют мультимножества переходов, относящиеся к нескольким изотопологам воды [15-17]), на практике применяются оба варианта декомпозиции – декомпозиции по согласованному и несогласованному мультимножествам переходов.

Представление результатов декомпозиции (E, F) в форме, допускающей разделение на расчетные и измеренные данные, позволяет качественно характеризовать доверие. В подавляющем большинстве случаев в количественной спектроскопии доверие к измеренным характеристикам переходов существенно больше, чем к вычисленным.

Приведем два примера оценки доверия экспертных массивов данных из [2,8]. Для изотополога воды $H_2{}^{18}O$ из 9753 переходов, содержащихся в экспертных данных, только 80% переходов совпадает с первичными данными измерений, 33%

совпадает с данными вычислений и 85% совпадает с данными измерений и расчетов. Для изотополога диоксида углерода $^{13}CO_2$ из 49777 экспертных переходов только 18% совпадает с измеренными данными и 99% совпадает с данными расчетов. Несмотря на то, что у изотополога диоксида углерода совпадение с первичными данными более полное, доверие к экспертному массиву данных для молекулы $H_2{}^{18}O$ может быть существенно большим.

4.3. Представления оценок доверия к информационным ресурсам

Результаты решения задачи D1 оценки доверия к информационным ресурсам в онтологии информационных ресурсов представлены индивидами класса DecompositionInformationSource (в этой работе используется терминология онтологии версии 6). Декомпозиция осуществляется тремя способами: по экспериментальным первичным данным, по теоретическим первичным данным и по экспериментальным и первичным данным. Этот принцип не зависит от того, по какой физической величине производится проверка критерия опубликования. Проверка критерия опубликования для вакуумных волновых чисел, выполненная в данной работе, использует разбиение интервала изменения волновых чисел (от 0 до 100000 $см^{-1}$) на 12 интервалов, с каждым из которых связана предельная точность, определяющая предел применимости критерия опубликования при проверке экспертного массива.

4.3.1. Табличное представление оценки доверия

Для полной оценки качества экспертных данных по критерию опубликования приведем результаты наших исследований на примере молекулы диоксида углерода. Таблица, отображенная на Рис.6, содержит результаты декомпозиции экспертных данных, относящихся к основному изотопологу молекулы диоксида углерода CO2. Она является часть описания свойств экспертного источника данных. В верхней части таблицы размещены интервалы изменения значений вакуумных волновых чисел молекулы. Имеющиеся в настоящее время данные относятся к интервалам от дальнего до ближнего инфракрасного диапазона. Декомпозиция проведена тремя способами: по теоретическим данным, по экспериментальным данным и по совокупности теоретических и экспериментальных данных. При разложении по первым двум способам предоставляется информация по каждому первичному источнику данных, с которым экспертные данные имеют идентичные переходы, удовлетворяющие критерию опубликованию и интервал изменения волновых чисел этих доверяемых чисел.

Декомпозиция экспертного источника данных													
Нотация квантовых чисел: HitranNotation								Просмотреть результаты декомпозиции					
Тип декомпозиции.	*Вакуумные волновые числа*												
	Радио	*Микро-вволновой*	*Дальний ИК*	*Длинно-волновой ИК*	*Средний ИК*	*Коротко-волновой ИК*	*Ближний ИК*	*Видимый*	*Ближний УФ*	*Средний УФ*	*Дальний УФ*	*Рентгеновский*	*Все диапазоны*
Декомпозиция по расчетным данным	Интервал [630.0, 1250.0). Точность сравнения 0.005. Единицы измерения [1/см].												
2003_TaPeTeBy_a_CO2			*436.123 629.983 [2730]*	*630.007 1110.727 [7905]*	*1801.972 3299.927 [9536]*	*3300.370 6988.655 [15439]*	*7248.072 8309.764 [1025]*						*436.123 8309.764 [36635]*
1965_GoMc_CO2			*585.718 628.931 [28]*	*630.482 754.655 [187]*									*585.718 754.655 [215]*
Сомнительные волновые числа			*593.844 612.242 [13]*	*649.453 801.156 [52]*	*1955.339 2396.667 [84]*	*3476.214 6851.044 [46]*	*7416.308 9531.635 [69]*						*593.844 9531.635 [264]*
Декомпозиция по данным измерений													
2007_Horneman_CO2			*593.660 628.931 [42]*	*630.482 749.841 [224]*									*593.660 749.841 [266]*
1983_JoKaHo_CO2			*584.123 629.233 [56]*	*630.482 752.832 [302]*									*584.123 752.832 [358]*
Сомнительные волновые числа			*352.066 629.991 [7592]*	*630.007 1197.483 [19286]*	*1418.392 3299.983 [25151]*	*3300.075 6999.862 [46782]*	*7000.026 12783.875 [9059]*						*352.066 12783.875 [107870]*
Декомпозиция по данным измерений и расчетов													
Сомнительные волновые числа			*593.844 612.242 [13]*	*649.453 801.156 [52]*	*1955.339 2396.667 [84]*	*3476.214 6821.572 [26]*	*7416.308 9531.301 [27]*						*593.844 9531.301 [202]*

Рис. 6. Табличное представление результатов декомпозиции экспертных данных по молекуле CO2, извлеченных из работы [8].

Размещенная в ячейки таблицы колонка чисел описывает нижнюю и верхнюю границы изменения доверительных волновых чисел в указанном интервале, а в квадратных скобках помещено число доверительных волновых чисел. Данное число является также гиперссылкой на таблицу, изображенную на рис. 7 в которой в явном виде показаны доверительные волновые и числа и разность их величин. Для каждого из способов декомпозиции в строках «Сомнительные волновые числа» представлены интервалы изменения волновых чисел и их число. Число в квадратных скобках также является гиперссылкой на таблицу, содержащую сомнительные волновые числа полученные при соответствующем способе декомпозиции.

2009_RoGoBaBe_CO2 **(Left) to (Right)** 2007_Horneman_CO2 **в интервале Дальний ИК [10.0 , 630.0) с точностью сравнения 0.005 . Единицы измерения [1/см].** Расчет/Эксперимент

Показать 5 строк от 0 — **Всего строк 42** — > >> — Настройки

N	Вакуумные волновые числа PQ ($см^{-1}$) Left	Вакуумные волновые числа PQ ($см^{-1}$) Right	$\lvert PQ_{Left} - PQ_{Right} \rvert$	$n_1^{up\ HN}$	$n_2^{up\ HN}$	$l_2^{up\ HN}$	$n_3^{up\ HN}$	$r^{up\ HN}$	$sym^{up\ HN}$	$j^{up\ HN}$	$n_1^{low\ HN}$	$n_2^{low\ HN}$	$l_2^{low\ HN}$	$n_3^{low\ HN}$	$r^{low\ HN}$	$sym^{low\ HN}$	$j^{low\ HN}$
1	593.660291	593.660309	0.000018	1	0	0	0	2	e	30	0	1	1	0	1	e	31
2	595.242846	595.242863	0.000017	1	0	0	0	2	e	28	0	1	1	0	1	e	29
3	596.823691	596.823707	0.000016	1	0	0	0	2	e	26	0	1	1	0	1	e	27
4	598.40291	598.402947	0.000037	1	0	0	0	2	e	24	0	1	1	0	1	e	25
5	599.98058	599.980581	0.000001	1	0	0	0	2	e	22	0	1	1	0	1	e	23

Показать 5 строк от 0 — **Всего строк 42** — > >> — Настройки

Рис.7. Табличное представление идентичных переходов, удовлетворяющих критерию опубликования в ближнем ИК диапазоне.

Таблица 1 содержит результаты декомпозиции экспертных данных из работы [8] для всех изотопологов этой молекулы. В ней результаты исследования представлены в колонках (от радиочастотного диапазона до видимого диапазона). Экспертные источники данных характеризуются аббревиатурой и библиографической ссылкой. В первой колонке указаны изотопологи воды. Последующие колонки содержат числа, описывающие проценты содержания в указанном диапазоне сомнительных волновых чисел. Отметим, что до сих пор не проведен анализ полноты для данных о изотопологов молекулы CO2 и не опубликованы результаты согласования согласование данных.

Таблица 1. Результаты декомпозиции экспертных данных [8] для 6 изотопологов диоксида углерода.

Интервал декомпозиции	Вакуумные волновые числа					
	Микроволновой	Дальний ИК	Длинноволновой ИК	Средний ИК	Коротковолновой ИК	Ближ ИК
$^{12}C^{16}O_2$	128170					
Сомнительные данные		13	53	84	26	27
$^{12}C^{16}O^{17}O$	19264					
Сомнительные данные		38	187	29	926	0
$^{12}C^{16}O^{18}O$	79958					
Сомнительные данные	7	79	944	214	433	0
$^{13}C^{16}O_2$	49777					
Сомнительные данные		0	0	81	41	1
$^{13}C^{16}O^{17}O$	2953					
Сомнительные данные		127	364	590	0	
$^{12}C^{17}O^{18}O$	821					
Сомнительные данные		6	59	90	0	

В серых полях таблицы указано число сомнительных переходов в соответствующем диапазоне. Числа напротив химических формул молекул соответствуют числу переходов в экспертных данных.

4.3.2. Онтологическое представление оценки доверия

Разбиение интервала изменения волновых чисел (от 0 до 100000 см$^{-1}$) на 12 интервалов приводит к необходимости создания 17 классов в онтологии, описывающей доверие к вакуумным волновым числам, содержащихся в экспертных информационных ресурсах. Четыре класса связаны со способами декомпозиции (пример класса DescriptionUnderExperimentalPrimaryInformationSource Decomposition) и тринадцать классов с интервалами декомпозиции (пример класса DecompositionInFarInfraredRangeDescription). Два класса TrustedDescription и DistrustedDescription содержат элементы, описывающие число доверяемых и сомнительных вакуумных волновых чисел и интервалы изменения этих волновых чисел.

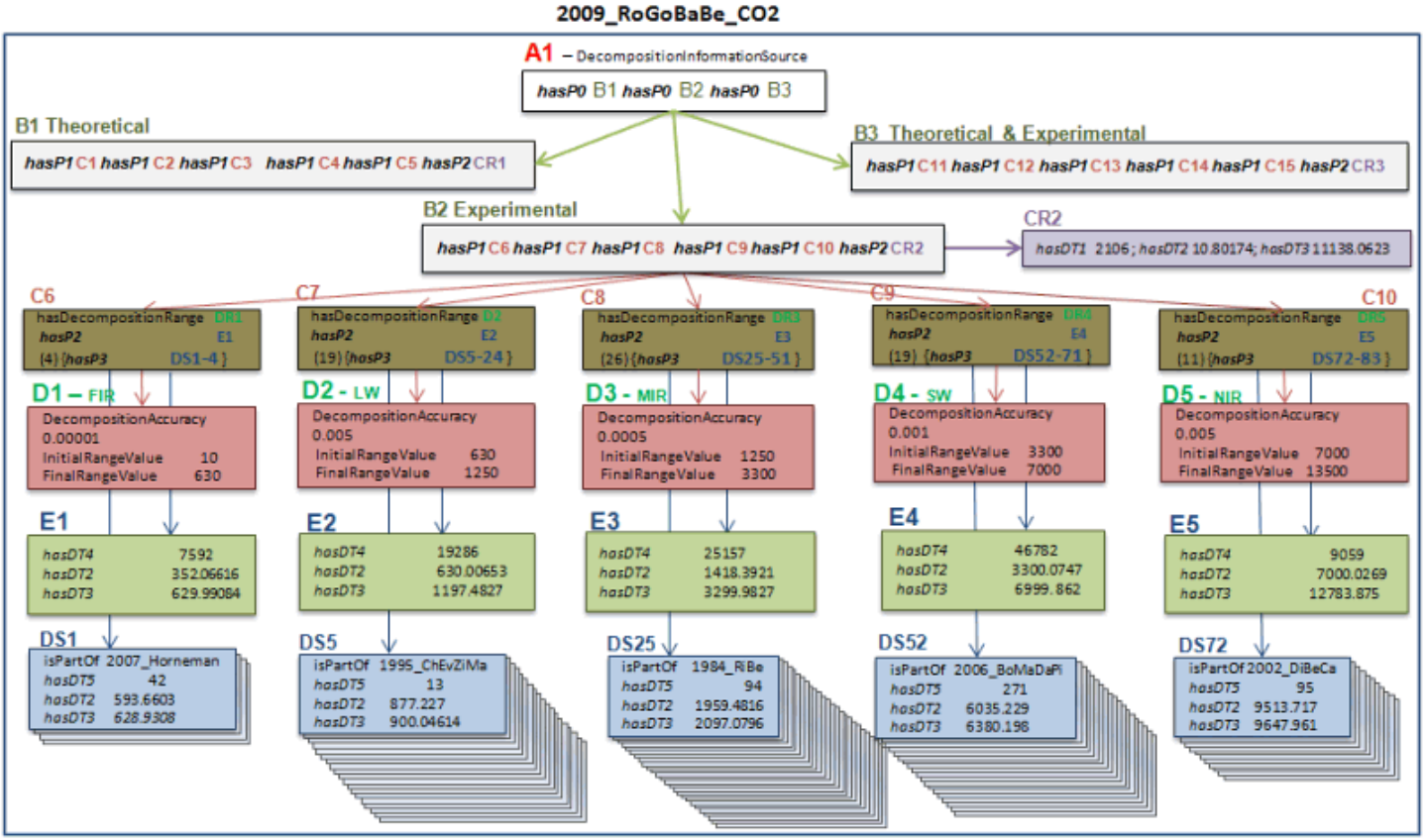


Рис.8. Структура индивида, характеризующего оценку доверия к экспертным данным по молекуле CO2.

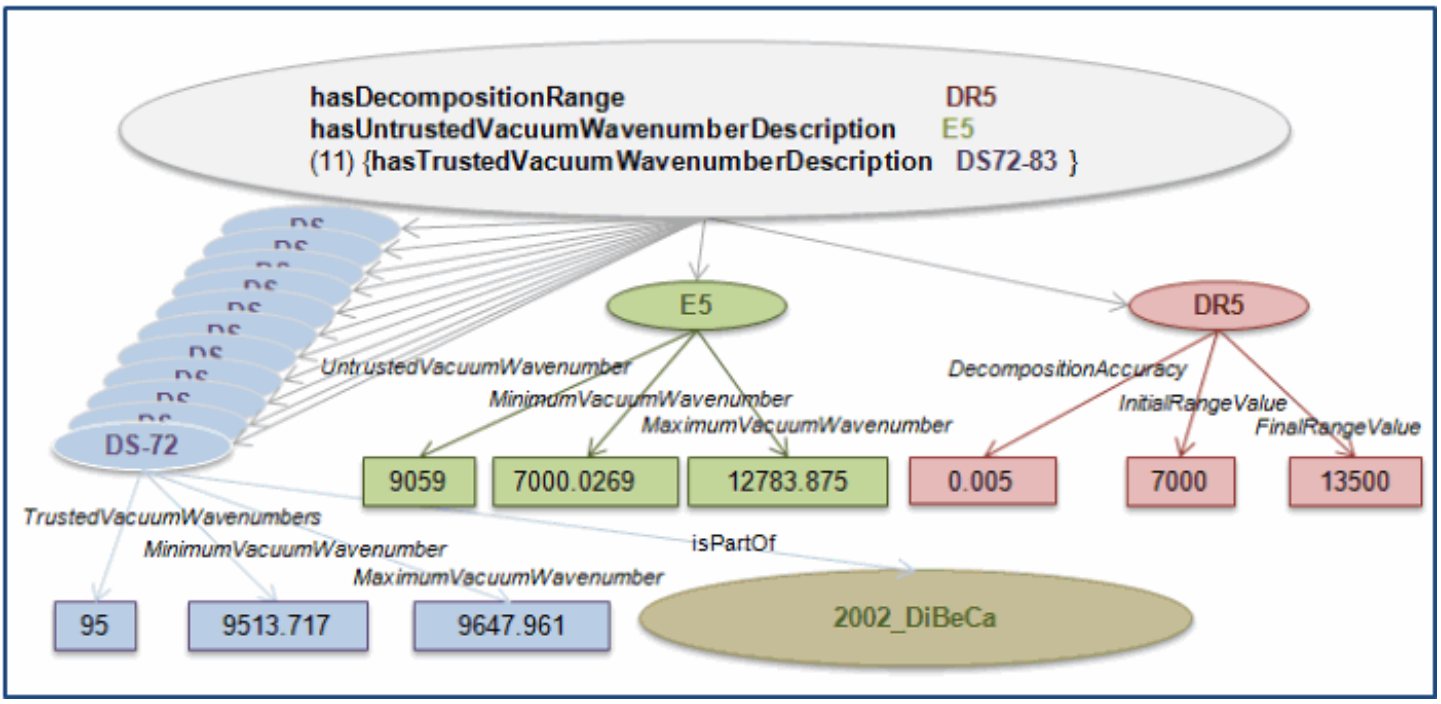


Рис. 9. Фрагмент рис.7, показывающий в деталях структуру индивида класса **DecompositionNearInfraredRangeDescription** (C10).

Структура индивида этого класса представлена на рис.9. Этот индивид обладает тремя объектными свойствами, значениями которых являются индивиды DR, E, DS, описывающие интервал декомпозиции, сомнительные и доверительные характеристики экспертных данных в этом диапазоне, соответственно. Доверительные характеристики отнесены к каждому из первичных источников данных, которые имеют идентичные переходы с экспертными данными, и волновые числа этих переходов удовлетворяют ограничению заданному критерием опубликования в этом диапазоне. На рис. 9 представлены свойства только одного из индивидов DS, а свойства индивида 2002_DiBeCa не показаны. Последний индивид является источником информации, описывающим свойства решения задачи T6, опубликованного в работе [24], используя при этом 100 аксиом.

Принимая во внимание остальные индивиды, характеризующие доверительные переходы экспертных данных, а также подобные индивиды, соответствующие другим интервалам декомпозиции и разложениям по разным комбинациям первичных источников данных получаем детальное описание доверительной части переходов по вакуумным волновым числам.

Рассмотрим вопрос, относящийся к качеству данных, связанный с критерием опубликования. Какие волновые числа в экспертных данных в ближнем инфракрасном диапазоне являются доверяемыми и какие волновые числа являются сомнительными при декомпозиции по измеренным и предсказанным данным? Класс **DecompositionNearInfraredRangeDescription** определен с помощью ограничений на значения свойств (используется манчестерский синтаксис для онтологии версии 6) и они приведены ниже.

(hasTrustedDescription **some** TrustedVacuumWavenumberDescription

or

hasUntrustedDescription **some** UntrustedVacuumWavenumberDescription)

and

(hasDecompositionRange value NearInfraredDecompositionRange)

Подобные ограничения используются для определения остальных 20 классов прикладной онтологии.

Онтология, описывающая оценку доверия к экспертным информационным ресурсам (см. http://wadis.saga.iao.ru/saga2/meta/get/V5_CO2_Expert_T6_DecompositionFor_global.owl), обладает выразительностью ALCHOIN(D).

5. Заключение

В работе исследовано качество экспертных спектральных ресурсов, относящихся к молекуле диоксида углерода. Исследование опирается на первичные данные, извлеченные из более чем 700 источников, размещенные в ИС W@DIS [13] и обеспеченные семантическими аннотациями. Аннотации содержат информацию о достоверности и степени согласования спектральных данных и доступны для пользователей в форме OWL-онтологий и через пользовательские интерфейсы.

Для оценки доверия к экспертным ресурсам используется критерий опубликования с ограничениями, количественные значения которых приведены для ряда интервалов изменения вакуумных волновых чисел. Описаны интерфейсы для просмотра пользователями части семантических аннотаций, представляющих пользователям информацию о доверяемых и сомнительных вакуумных волновых числах. Оценка доверия к экспертным ресурсам выполнена для экспертных значений вакуумных волновых чисел [8]. Показано, что подавляющее число волновых чисел из экспертных данных по молекуле диоксида углерода и ее изотопологов взято из расчетных данных для основного изотополога число сомнительных данных составляет 10^{-3}%. Число сомнительных данных для других изотопологов существенно выше ($^{12}C^{16}O^{17}O$ – 6%, $^{12}C^{16}O^{18}O$ – 2%, $^{13}C^{16}O_2$ – 0.2%, $^{13}C^{16}O^{17}O$ – 37% и $_{12}C_{17}O_{18}O$ – 18%).

Планируется применить вычисленную онтологическую базу знаний для построения экспертной системы для оценки информационных ресурсов и семантического поиска информационных ресурсов в количественной спектроскопии и Виртуальном центре атомных и молекулярных данных [5]. Такая база знаний, с одной стороны, позволяет проводить анализ сторонних для нас экспертных данных по ряду молекул, а с другой - формировать для экспертов согласованные с опубликованными данными экспертные данные, предоставляя экспертам количественные оценки достоверности и доверия таких данных.

Данная статья является пятой частью серии статей об онтологической базе знаний в молекулярной спектроскопии [14, 24, 25, 37].

Авторы благодарны Российскому фонду фундаментальных исследований за финансовую поддержку работы (гранты 11-07-00660-а и 13-07-00411) и Лаврентьеву Ф.А. за помощь в подготовке и импорте опубликованных первичных данных, содержащих спектральные характеристики молекулы диоксида углерода и ее изотопологов.

Литература

1. Richard C., New section of the HITRAN database: Collision-induced absorption (CIA) / Richard C., Gordon I.E., Rothman L.S., Abel M., Frommhold L., Gustafsson M., Hartmann J.-M., Hermans C., Lafferty W.J., Orton G.S., Smith K.M., Tran H. // Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 113, Issue 11, July 2012, Pages 1276–1285.

2. Jacquinet-Husson N., The 2009 edition of the GEISA spectroscopic database / Jacquinet-Husson N., Crepeau L., Armante R., Boutammine C., Chedin A., Scott N.A., Crevoisier C., Capelle V., Boone C., Poulet-Crovisier N., Barbe A., Campargue A., Chris Benner D., Benilan Y., Bezard B., Boudon V., Brown L.R., Coudert L.H., Coustenis A., Dana V., Devi V.M., Fally S., Fayt A., Flaud J.-M., Goldman A., Herman M., Harris G.J., Jacquemart D., Jolly A., Kleiner I., Kleinbohl A., Kwabia-Tchana F., Lavrentieva N., Lacome N., Li-Hong Xu, Lyulin O.M., Mandin J.-Y., Maki A., Mikhailenko S., Miller C.E., Mishina T., Moazzen-Ahmadi N., Muller H.S.P., Nikitin A., Orphal J., Perevalov V., Perrin A., Petkie D.T., Predoi-Cross A., Rinsland C.P., Remedios J.J., Rotger M., Smith M.A.H., Sung K., Tennyson J., Toth R.A., Vandaele A.-C., J. Vander Auwera // Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 112, Issue 15, October 2011, Pages 2395-2445.

3. Chance K., An improved high-resolution solar reference spectrum for earth's atmosphere measurements in the ultraviolet, visible, and near infrared / Chance K., Kurucz R.L. // Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 111, Issue 9, June 2010, Pages 1289–1295.

4. Hase F., The ACE-FTS atlas of the infrared solar spectrum / Hase F., Wallace L., McLeod S.D., Harrison J.J., Bernath P.F. // Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 111, Issue 4, March 2010, Pages 521–528.

5. Dubernet M.L., Virtual atomic and molecular data centre / Dubernet M.L., Boudon V., Culhane J.L., Dimitrijevic M.S., Fazliev A.Z., Joblin C., Kupka F., Leto G., Le Sidaner P., Loboda P.A., Mason H.E., Mason N.J., Mendoza C., Mulas G., Millar T.J., Nunez L.A., Perevalov V.I., Piskunov N., Ralchenko Y., Rixon G., Rothman L.S., Roueff E., Ryabchikova T.A., Ryabtsev A., Sahal-Brechot S., Schmitt B., Schlemmer S., Tennyson J., Tyuterev V.G., Walton N.A., Wakelam V. and Zeippen C.J. // Journal of Quantitative Spectroscopy and Radiative Transfer, 2010, Volume 111, Issue 15, Pages 2151-2159.

6. Cami, J., SPECTRAFACTORY.NET: A database of molecular model spectra / Van Malderen, R., Markwick, A.J. // Astrophysical Journal, Supplement Series, Volume 187, Issue 2, 2010, Pages 409-415 DOI: 10.1088/0067-0049/187/2/409.

7. Rothman L.S., HITEMP, the high-temperature molecular spectroscopic database / Rothman L.S., Gordon I.E., Barber R.J., Dothe H., Gamache R.R., Goldman A., Perevalov

V.I., Tashkun S.A., Tennyson J. // Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 111, Issue 15, October 2010, Pages 2139-2150.

8. Rothman L.S., The HITRAN 2008 molecular spectroscopic database / Gordon I.E., Barbe A., Benner D.Chris, Bernath P.F., Birk M., Boudon V., Brown L.R., Campargue A., Champion J.-P., Chance K., Coudert L.H., Dana V., Devi V.M., Fally S., Flaud J.-M., Gamache R.R., Goldman A., Jacquemart D., I. // Journal of Quantitative Spectroscopy and Radiation Transfer, 2009, Volume 110, Issue 9, Pages 533-572.

9. Toth R.A., Spectroscopic database of CO2 line parameters: 4300–7000 cm-1 / Toth R.A., Brown L.R., Miller C.E., Malathy Devi V. and Benner D.Chris // Journal of Quantitative Spectroscopy and Radiation Transfer, 2008, Volume 109, Issue 6, Pages 906-921.

10. Tran H., Tran H., Model, software and database for line-mixing effects in the ?3 and ?4 bands of CH4 and tests using laboratory and planetary measurements—II: H2 (and He) broadening and the atmospheres of Jupiter and Saturn / Tran H., Flaud P.-M., Fouchet T., Gabard T., Hartmann J.-M. // Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 101, Issue 2, September 2006, Pages 306–324.

11. Tran H., Model, software and database for line-mixing effects in the ?3 and ?4 bands of CH4 and tests using laboratory and planetary measurements—I: N2 (and air) broadenings and the earth atmosphere / Tran H., Flaud P.-M., Gabard T., Hase F., von Clarmann T., Camy-Peyret C., Payan S., Hartmann J.-M., , Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 101, Issue 2, September 2006, Pages 284–305.

12. Лаврентьев Н.А., Сравнение спектральных массивов данных HITRAN и GEISA с учетом ограничения на опубликование спектральных данных / Лаврентьев Н.А., Макогон М.М., Фазлиев А.З. // Оптика атм. и океана. 2011. Т.24. №4. С.279-292.

13. Информационная система W@DIS, [Электронный ресурс]. – Режим доступа: http://wadis.saga.iao.ru

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

15. Tennyson J., IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part I. Energy Levels and Transition Wavenumbers for H217O and H218O / Tennyson J., Bernath P.F., Brown L.R., Campargue A., Carleer M. R., Csaszar A.G., Gamache R.R., Hodges J.T., Jenouvrier A., Naumenko O.V., Polyansky O.L., Rothman L.S., Toth R.A., Vandaele A.C., Zobov N.F., Daumont L., Furtenbacher T., Fazliev A., Gordon I.E., Mikhailenko S.N., Shirin S.V. // Journal of Quantitative Spectroscopy and Radiative Transfer, 2009, v.110, no.9-10, p.573-596.

16. Tennyson J., IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part II. Energy Levels and Transition Wavenumbers for HD16O, HD17O, and

HD18O / Tennyson J., Bernath P.F., Brown L.R., Campargue A., Carleer M. R., Csaszar A.G., Daumont L., Gamache R.R., Hodges J.T., Naumenko O.V., Polyansky O.L., Rothman L.S., Toth R.A., Vandaele A.C., Zobov N.F., Fall S., Furtenbacher T., Fazliev A., Gordon I.E., Shui-Ming Hu, Mikhailenko S.N., Voronin B.A. // Journal of Quantitative Spectroscopy and Radiative Transfer, 2010, v.111, no.15, p. 2160-2184.

17. Tennyson J., IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part III. Energy Levels and Transition Wavenumbers for H216O / Tennyson J., Bernath P.F., Brown L.R., Campargue A., Csaszar A.G., Daumont L., Gamache R.R., Hodges J.T., Naumenko O.V., Polyansky O.L., Rothman L.S., Vandaele A.C., Zobov N.F., Al Derzi A.R., Fabrie C., Fazliev A., Furtenbacher T., Gordon I.E., Lodi L., Mizus I. // Journal of Quantitative Spectroscopy and Radiative Transfer, 2013, v.117, no., p.29-58.

18. Половцева Е.Р., Информационная система для решения задач молекулярной спектроскопии. 5. Колебательно-вращательные переходы и уровни энергии молекулы H2S / Половцева Е.Р., Лаврентьев Н.А., Воронина С.С., Науменко О.В., Фазлиев А.З. // Оптика атм. и океана. 2011, Т.24, №10, с. 898-905.

19. Tashkun S.A., Critical evaluation of measured pure-rotation and rotation-vibration line positions and an experimental dataset of energy levels of 12C16O in X1?+ state / Tashkun S.A., Velichko T.I. and Mikhailenko S.N. // Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 111, Issue 9, June 2010, Pages 1106-1116,

20. Voronina S.S., Systematization of the published spectroscopic parameters of ammonia / Voronina S.S., Yurchenko S.N. and Fazliev A.Z. // Abstracts of the 22-nd Colloquium on High Resolution Molecular Spectroscopy, 2011, p.163.

21. Lavrentiev N.A., Complete set of published spectral data on CO2 molecule / Lavrentiev N.A., Privesentsev A.I., Filippov N.N., and Fazliev A.Z. // Abstracts of the 22-nd Colloquium on High Resolution Molecular Spectroscopy, Dijon, 2011, p.353.

22. Voronina S.S., Systematization of Published Spectral Data on N2O and OCS molecules and Isotopologues / Voronina S.S., Privezentsev A.I., and Fazliev A.Z.,. Abstracts of the 23-nd Colloquium on High Resolution Molecular Spectroscopy, Budapest, 2013.

23. Akhlestin A.Y., Systematization of published data on phosphine isotopologoues / Akhlestin A.Y., Voronina S.S., Privesentsev A.I., Fazliev A.Z. // Proc. of SPIE. 18-th Inter. Symp. on Atmos. and Ocean Optics. Atmos. Physics. Irkutsk, Rus. Federation, 2-6 July 2012. V.8696. P.8696-38.

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

25. Ахлёстин А.Ю., Базы знаний для описания информационных ресурсов в молекулярной спектроскопии 4. Программное обеспечение / Ахлёстин А.Ю.,

Козодоев А.В., Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З., , Электронные библиотеки, 2012, т. 15, в.3. [Электронный ресурс]. – Режим доступа: http://elbib.ru/index.phtml? page=elbib/ rus/journal/2012/part3/AKLPF

26. Bykov A.D., Distributed information system on atmospheric spectroscopy / Bykov A.D., Fazliev A.Z., Filippov N.N., Kozodoev A.V., Privezentsev A.I., Sinitsa L.N., Tonkov M.V., Tretyakov M.Yu. // Geophysical Research Abstracts, Vol. 9, 01906, 2007 SRef-ID: 1607-7962/gra/EGU2007-A-01906.

27. De Roure D., A Future e-Science Infrastructure / De Roure D., Jennings N., Shadbolt N. // Report commissioned for EPSRC/DTI Core e-Science Programme. 2001. 78p.

28. Ахлёстин А.Ю., Информационная система трехслойной архитектуры / Ахлёстин А.Ю., Лаврентьев Н.А., Привезенцев А.И., Фазлиев А.З. // Труды семинара «Наукоемкое программное обеспечение» 2011, Новосибирск, с.38-43.

29. Crolek T. Matthew, The Six Quests for the Electronic Grail: Current Approaches to Information Quality in WWW Resources // Extrait de la Revue Informatique et Statistique dans les Sciences humaines, XXXII, 1 a 4, 1996. C.I.P.L. - Universite de Liege - Tous droits reserves.

30. Bizer C., Using Context- and Content-Based Trust Policies on the Semantic Web / Bizer C., Oldakowski R. // WWW 2004, May 17-22, 2004, New York, NY USA.

31. O'Hara K., Trust Strategies for the Semantic Web / O'Hara K., Alani H., Kalfoglou Y. and Shadbolt N. // Workshop on Trust, Security, and Reputation on the Semantic Web, 3rd International (ISWC'04), Hiroshima, Japan, 07 - 11 Nov 2004. (2004).

32. Sabater J. Review on Computational Trust and Reputation Models / Sabater J. and Sierra C. // Artificial Intelligence Review, (2005) v.24, p.33–60.

33. Artz D., A survey of trust in computer science and the Semantic Web / Artz D., Gil Y. // Web Semantics: Science, Services and Agents, on the World Wide Web,(2007) 58–71.

34. Gil Y., Towards content trust of web resources / Gil Y., Artz D. // Web Semantics: Science, Services and Agents on the World Wide Web, (2007) 227–239.

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

36. Апанович З.В., Цифровая библиотека научных статей по количественной спектроскопии / Апанович З.В., Винокуров П.С., Ахлёстин А.Ю., Привезенцев А.И., Фазлиев А.З. // Труды 14ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2012, Переславль, с.53-59.

37. Привезенцев А.И., Базы знаний для описания информационных ресурсов в молекулярной спектроскопии. 1. Основные понятия / Привезенцев А.И., Фазлиев А.З. // Электронные библиотеки, т.14, в.1, 2011
http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2011/part1

Приложение 1. Алгоритмы расчета показателей (схожести) разупорядочения A00 и A01.

Показатели разупорядочения количественно описывают разупорядочение. Таких показателей в данной работе три. Каждый из них связан с задачей упорядочения физических величин (в спектроскопии такими величинами могут быть вакуумные волновые числа, интенсивности и т.д., которые связаны с переходами), относящихся к состояниям или переходам физической системы. Значения этих величин измеряются с некоторой точностью, а переходы и состояния идентифицируются наборами чисел (квантовых чисел в спектроскопии).

Рассмотрим два спектральных массива (X,Y), содержащих уникальные идентичные переходы, характеризуемые значениями волновых чисел и квантовыми числами. Упорядочим массивы X и Y по значениям волновых чисел. Одинаково идентифицированные переходы массивов X и Y, в силу ошибок измерений и некорректного приписывания квантовых чисел, могут располагаться в разном порядке. Такая ситуация называется относительным разупорядочением. Она возникает при сравнении массивов содержащих идентичные величины.

Выше введены показатели разупорядочения A_{00}, A_{10} и A_{01}. Здесь сформулированы задачи по вычислению показателей разупорядочения и приведены алгоритмы их решения.

Пусть даны два множества X, Y элементами которых являются переходы. В каждом множестве все переходы уникальны и каждый элемент описывается значением физической величины и квантовыми числами (x_i, k_i) и (y_i, k_i). Пусть элементы X перенумерованы в порядке возрастания x и элементам приписаны порядковые номера (x_i, k_i, i). В элементах множества Y зафиксируем номера номер квантовых чисел соответствующий номеру квантовых чисел в последовательности (x_i, k_i, i) и припишем к элементам Y порядковый номер (y_j, k^i_j, j). Т.к. в последовательностях могут быть сделаны ошибки измерений или некорректные отнесения, то порядки следования квантовых чисел в этих последовательностях могут быть разными.

Отсюда следуют две задачи.

Задача 1. Найти минимальное число перестановок (A_{00}) в последовательности элементов множества Y в результате которых в последовательностях (x_i, k_i, i) и (y_j, k_j, j) порядок следования k_i, и k_j будет одинаковым.

Задача 2. Пусть последовательно, начиная с первого элемента упорядоченной последовательности множества X, сравниваются квантовые числа упорядоченных по значению физических величин множеств X и Y. Если квантовые числа первых

двух членов последовательности не совпадают, то их обеих последовательностей удаляют первый член последовательности Y и ему идентичный переход в X. Процедура удаления повторяется до тех пор, пока квантовые числа переходов не совпадут. При совпадении квантовых чисел совпавшие члены последовательностей оставляются в последовательностях, и подобная процедура проводится со вторым членом оставшейся последовательности элементов множества X. Найти число минимальное число (A_{01}) таких удалений которое необходимо сделать в последовательностях X и Y, чтобы члены последовательностей с одинаковыми номерами содержали одинаковые квантовые числа?

Отметим, что показатели разупорядочения неотрицательны и из $A_{00} = 0$ следует $A_{01}=0$ (обратное не верно).

Описание приведенных ниже алгоритмов дано в псевдосинтаксисе языка скриптов PHP.

Алгоритм решения задачи 1.

#Массив сначала отсортированный по физической величине wY с порядковым числом при сортировке idY,

#а затем отсортированный по физической величине wX с порядковым числом при сортировке idX

rowsXY = array(idX1:(idX1,idY1,qnsX1,wX1,wY1), ... ,

idXn:(idXn,idYn,qnsXn,wXn,wYn));

#Число элементов массива rowsXY,/p>

n=count(rowsXY); A00=0;

#Цикл прохода по всем элементам массива rowsXY

for(i=1;i<=n;i++) {(idXi,idYi,qnsXi,wXi,wYi) = rowsXY[i];

#Установить диапазон просмотра перестановок для текущей пары

#Диапазон просмотра перестановок вперед

if(idXi < idYi) {start=idXi;stop=idYi;

#Диапазон просмотра перестановок назад

} elseif(idXi > idYi) {start=idYi;stop=idXi-1;} else {continue;}

#Цикл прохода по диапазону перестановок

for(j=start+1;j<=stop;j++) {(idXj,idYj,qnsXj,wXj,wYj) = rowsXY[j];

#Диапазон просмотра перестановок вперед

if(idXi < idYi) {if(idYi > idYj) {A00+=1;}

#Диапазон просмотра перестановок назад

} elseif(idXi > idYi) {if(idYi < idYj) {A00+=1;}}}}

Алгоритм решения задачи 2.

#Ассоциативный массив отсортированный по физической величине wX

rowsX = array(1:(qnsX1,wX1), ... , n:(qnsXn,wXn));

#Ассоциативный массив отсортированный по физической величине wY

rowsY = array(1:(qnsY1,wY1), ... , n:(qnsYn,wYn));

#Число элементов массива rowsX

n=count(rowsX);

#Ассоциативный массив исключённых индексов из массива rowsX

exX = array();

#Ассоциативный массив исключённых индексов из массива rowsY

exY = array();

#Ассоциативный массив исключённых переходов из первого источника информации идентичный исключенным переходам второго источника

D = array();

#Цикл прохода по всем элементам массива rowsX

for(i=1;i<=n;i++) {

#Если индекс исключен, то перейти на следующую итерацию цикла rowsX

if (in_array(i, exX)) {continue;}(qnsXi, wXi) = rowsX[i];

#Установить индекс просмотра массива rowsY от текущего индекса просмотра массива rowsX

j=i;

```
#Найти следующий, не исключенный индекс массива rowsY

while(in_array(j, exY)) {j+=1;}

#Цикл прохода по всем элементам массива rowsY от j и до конца массива

while(j<=n) {(qnsYj, wYj) = rowsY[j];

#Если квантовые числа совпадают, то переход к следующей итерации

if (qnsXi==qnsYj){exY[] = j;break;} else {

#Установить индекс просмотра вложенного цикла по массиву rowsX от следующего значения главного индекса просмотра массива rowsX

k=i+1;

#Вложенный цикл прохода по всем элементам массива rowsX от k и до конца массива

while(k<=n) {

#Если индекс исключен, то перейти на следующую итерацию вложенного цикла rowsX

if (in_array(k, exX)) {k+=1;continue;} (qnsXk, wXk) = rowsX[k];

#Если квантовые числа совпадают, то заносим индексы в списки исключения и записываем найденную пару в массив исключенных переходов

if (qnsXk==qnsYj){exX[] = k;exY[] = j;D[k] = (qnsXk, (wXk, wYj));break;} else {k+=1;}}

j+=1;}}}

A01 = count(D)
```

Об авторах

Ахлёстин Алексей Юрьевич – программист Центра интергированных информационных систем Института оптики атмосферы им. В.Зуева СО РАН, e-mail: lexa@iao.ru

Лаврентьев Николай Александрович – научный сотрудник Центра интергированных информационных систем Института оптики атмосферы им. В.Зуева СО РАН, e-mail: lnick@iao.ru

Привезенцев Алексей Иванович – к.т.н., научный сотрудник Центра

интегрированных информационных систем Института оптики атмосферы им. В.Зуева СО РАН, e-mail: remake@iao.ru

Фазлиев Александр Зарипович – к.ф.-м.н., заведующий Центром интегрированных информационных систем Института оптики атмосферы им. В.Зуева СО РАН.