

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ
ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г.

№ 9

Москва 2015

К юбилею Юджина Гарфилда

16 сентября этого года выдающемуся представителю информационной профессии Юджину Гарфилду исполняется 90 лет. Его заслуги перед информационной наукой и практикой неocenимы. Он разработал новый способ раскрытия накопленных человечеством знаний, получивший название *системы цитирования*; создал ряд уникальных информационных продуктов – указатели содержания текущих журналов (*Current Contents*), библиографических ссылок в научной литературе (*Science Citation Index*), ссылок на журналы (*Journal Citation Reports*). Для реализации своих идей основал в г. Филадельфия (США, 1956 г.) уникальный информационный центр *Institute for Scientific Information*, который теперь принадлежит корпорации *Thomson Reuter* и является ядром системы *Web of Knowledge*.

Ю. Гарфилд – неумолимый ученый-исследователь, которого глубоко интересуют современные проблемы информатики и научной коммуникации; в течение более 20 лет еженедельно он публиковал свои статьи в выпусках *Current Contents*, которые составили 17 томов сборников *Essays of an Information Scientist* (1977 – 1994) – своего рода энциклопедию информатики.

В 1948 г. Ю. Гарфилд получил степень бакалавра по химии, в 1954 г. – степень магистра в Колумбийском университете (Нью-Йорк), в 1962 г. стал доктором наук по структурной лингвистике в Пенсильванском университете (Филадельфия), защитив диссертацию «Алгоритм перевода химических наименований в молекулярные формулы». Его работы по индексированию медицинской литературы (1951 г.) в Университете Джона Хопкинса (Балтимор, шт. Мэриленд) заложили теоретическую и практическую основу создания известной системы анализа и поиска этой литературы (*MEDLARS*). В 1956 г. он создал фирму, позднее названную Институтом научной информации (по его признанию, не без влияния прежнего названия ВИНТИ РАН), просуществовавшую до 1992 г., когда ее купил канадский газетный магнат Томсон.

С давних пор Ю. Гарфилд друг нашей страны, он любил приезжать к нам для чтения лекций, на книжные ярмарки.

Ю. Гарфилд является иностранным членом Российской академии естественных наук (1996 г.), ему присвоено звание почетного доктора наук Дальневосточного государственного университета (1997 г.).

Идеи Ю. Гарфилда, созданные им информационные системы и издания сегодня занимают одно из ведущих мест в информационном обеспечении науки, промышленности, экономики и культуры. Государственные органы и владельцы частных фирм оценивают эффективность своих инвестиций по библиографическим ссылкам на статьи, в которых излагаются результаты исследований, проводившихся на вложенные ими средства. Крупнейшие информационные системы *Web of Science* (*Thomson Reuters*), *Scopus* (*Elsevier*), *Springer*, *EBSCO*, Российский индекс научного цитирования отслеживают потоки литературы при помощи инструментов Ю. Гарфилда.

Редакция и редколлегия нашего журнала, наши авторы и читатели, сотрудники Всероссийского института научной и технической информации Российской академии наук сердечно поздравляют Вас, дорогой Джин. Мы счастливы, что Вы в такой великолепной форме и продолжаете работать на благо нашего дела. Долгих Вам лет жизни, здоровья и успехов.

Д.Л. Олышанский

Информационная среда WebChem-JSM для реализации ДСМ-метода в задачах фармакологии

Описывается программный комплекс, предоставляющий информационную среду для применения ДСМ-метода в задачах выявления связи структура-активность в фармакологии и смежных областях. Для решения поставленных задач предлагается клиент-серверная архитектура, с применением веб-технологий. По сравнению с предшествующими решениями достигается централизованное хранение данных и возможность наполнения базы знаний. Модульная архитектура комплекса создает богатые возможности для применения различных версий компонент, сравнения решателей и стратегий ДСМ-метода. Впервые для большей эффективности в интеллектуальной ДСМ-системе применяется параллельный решатель.

Ключевые слова: ДСМ-метод, информационная среда, веб-интерфейс, код ФКСП, биологическая активность

ВВЕДЕНИЕ

Установление связи между структурой и биологической активностью химических соединений является важнейшей задачей в фармакологии. Решение этой задачи на ЭВМ основано на применении методов машинного обучения, строящих модели для предсказания активности по набору известных пар структура-активность. Собираательно такие модели называются QSAR (Quantative Structure Activity Relationship) или SAR (Structure Activity Relationship), когда речь идет о лишь качественном определении наличия или отсутствия свойства. Разработка новых моделей и применение существующих методов является областью интенсивных исследований [1, 2].

Среди прочих моделей машинного обучения ДСМ-метод привлекателен своей уникальной особенностью – явное получение гипотез (правил), которыми пользуется алгоритм классификации. Это позволяет подключить экспертов к отбору гипотез в базе знаний (БЗ), таким образом, ДСМ-метод является инструментом для получения новых знаний (Knowledge Discovery).

В связи с трудоемкостью манипуляций с естественным представлением данных – структурных формул, т.е. графов, применяются проблемно-ориентированные дескрипторные языки. В таком языке молекула описывается как набор отдельных подструктур или центров, а также взаимодействий между ними в виде кодовых слов (дескрипторов). Дескрипторное представление химических соединений повышает скорость обработки данных и их информативность. Среди различных дескрипторных языков код ФКСП (фрагментарный код суперпозиций подструктур) хо-

рошо зарекомендовал себя для применения в задачах предсказания биологической активности интеллектуальными системами типа ДСМ. ФКСП приписывает соединению строку (код), описывающую отдельные активные подструктуры трех типов: линейные, циклические и замещения в ароматических группах. Недостатком такого представления является рассыпание молекул – невозможно однозначно описать расположение подструктур по коду [3-5].

Введем основные понятия, которые будем использовать в дальнейшем. ДСМ-эксперимент или просто эксперимент – выполнение задачи классификации средствами ДСМ-метода. Основа эксперимента – это набор пар из соединений и приписанных им активностей в дискретной шкале. Эти пары мы будем называть родителями (гипотез) или примерами. Для ДСМ-метода примеры будут описываться строкой кодов ФКСП (атрибуты) и строкой активности (свойства). Гипотеза – обобщение нескольких примеров, полученное операцией сходства – пересечением их атрибутов и свойств. t-примерами (тау-примерами) называются примеры, активность которых необходимо доопределить в ходе классификации после генерации гипотез.

Главные задачи разрабатываемого программного комплекса:

- создать информационную среду для проведения ДСМ-экспериментов в области фармакологии;
- проводить автоматическое кодирование химических соединений в дескрипторный язык;
- вести базу данных (БД) по проводимым ДСМ-экспериментам;

- формировать базу знаний (БЗ) с участием экспертов по результатам экспериментов;
- предоставить пользователям визуализацию данных предметной области, гипотез и доопределений.

Желательно обойти проблемы прежних систем в этой области: монолитность, зависимость от среды выполнения. Необходимо построить модульный комплекс, создать удобную оболочку (среду) для главных компонентов. Также необходимо сформировать список основных требований, которыми должна руководствоваться реализация:

1. Удобство применения и доступность. Система не должна требовать установки на компьютер пользователя, требовать настройки параметров, необходимо избежать зависимости от операционной системы (ОС). Использование системы с различных устройств не должно вызывать дополнительных трудностей.

2. Централизованное хранение данных в БД и БЗ. При таком подходе мы избежим потерь и размывания результатов исследований по различным частным хранилищам.

3. Поддержка многих одновременных пользователей. При этом должно соблюдаться разграничение прав доступа, где каждый пользователь может работать лишь со своими данными.

4. Гостевой доступ. В качестве расширения многопользовательского режима должна быть возможность предоставить гостевой или временный доступ к системе отдельным пользователем. Также нужно иметь возможность контролировать объем ресурса, используемого пользователями.

5. Высокая эффективность и масштабирование. Все входящие в комплекс компоненты должны быть рассчитаны на высокую производительность, эффективно использовать ресурсы современных ЭВМ. При этом комплекс должен предусмотреть расширение вычислительных мощностей для больших задач без существенных изменений в его программной архитектуре.

6. Гибкость. В любую систему необходимо вносить изменения и комплекс должен иметь простую процедуру обновления программного обеспечения (ПО), а также замены отдельных компонент, расширения возможностей.

Пользователи информационной среды делятся на две категории – эксперты предметной области (химия) и разработчики интеллектуальных ДСМ-систем. Интересы первой категории пользователей, это удобный интерфейс, высокая скорость работы и минимальная сложность в настройке системы. Вторые заинтересованы в исследовании новых стратегий генерации гипотез, новых выборках данных для тестов, применении и сравнении различных реализаций ДСМ-метода. Важно создать платформу для гармоничного сотрудничества этих групп, при которой они не мешают работе друг друга, а плодотворно применяют достижения коллег в своих исследованиях.

АРХИТЕКТУРА ПРОГРАММНОГО КОМПЛЕКСА

Необходимость централизованно хранить информацию практически сразу приводит нас к клиент-серверной архитектуре, где множество клиентских

программ взаимодействуют с (логически) единственной программой-сервером. В таком сценарии БД и БЗ будут храниться на стороне сервера (т.е. доступны только ему), а пользователи будут взаимодействовать с этими базами лишь через регламентированный интерфейс. Клиент-серверная архитектура является хорошо изученной схемой разделения обязанностей в сложных системах. В нашем случае она дает ряд желаемых качеств – поддержка множества пользователей и гибкость, так как обновление ПО сервера (при условии соблюдения протоколов) может происходить без какого-либо участия пользователей.

Таким образом, следующим важным аспектом становится выбор протокола взаимодействия и технологий, на которых будут построены клиентская и серверная части комплекса. Следуя требованиям максимально упростить или ликвидировать установку дополнительных программ на компьютеры пользователей, практически единственным доступным выбором для реализации программы-клиента становится веб-браузер. Влияние веб-технологий и Интернета на вычислительные системы сложно переоценить и веб-браузер можно обнаружить в составе практически любой современной ОС. Протокол связи в этом случае, соответственно – HTTP, а программа-клиент представляется самим сервером в виде набора HTML – страниц, стилей CSS и программ на JavaScript. Это удачным образом решает проблему обновления ПО клиента – оно всегда оказывается той же версии, что и сервер. Еще одним преимуществом веб-браузера как платформы для клиента является удобство использования и возможность предоставить гостевой доступ. Применение стандартного протокола делает комплекс доступным практически из любой точки земного шара, где есть подключение к Интернету. Действительно, концепция веб-интерфейса применяется повсеместно, в том числе и для доступа к химическим базам и системам SAR [6].

При всем удобстве веб-браузера как клиента, отметим и неизбежные ограничения этой технологии, которые она накладывает на реализацию комплекса. Первое – это применение языка разметки HTML и стилей CSS для отображения пользовательского интерфейса, плохо пригодные для сложной визуализации, например, молекул. Вторым ограничением является модель запрос-ответ протокола HTTP, в этой схеме сервер не может сообщить о завершении какой-либо длительной процедуры без запроса клиента. Тем не менее, оба эти ограничения хорошо известны и имеют решения. В первом случае элемент Canvas новой спецификации HTML5 [7] дает богатые возможности по визуализации, для второй проблемы возможны несколько решений, но мы используем самое простое – периодический опрос состояния задачи на сервере.

Заключительным аспектом в архитектуре системы является внутренняя структура сервера – компоненты и их связи. Обзорная схема архитектуры комплекса приведена на рис. 1. Далее мы опишем назначение, ключевые особенности и критерии выбора технологий для каждого из функциональных блоков.

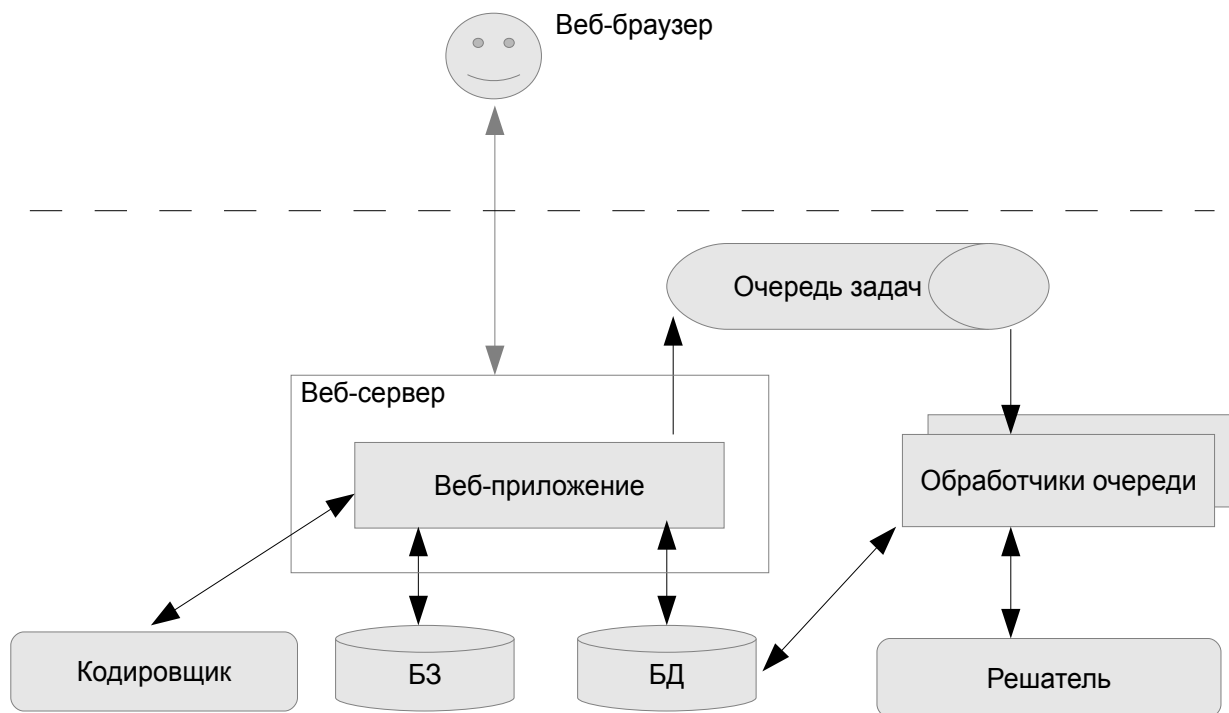


Рис.1. Модульная архитектура комплекса

Основные задачи веб-сервера – передача пользователю по протоколу HTTP статических файлов и JavaScript-программ для веб-браузера. Также, более существенно, реализуется базовый функционал HTTP для веб-приложения, которое непосредственно реализует лишь саму логику обработки запросов. Таким образом, веб-сервер является платформой для веб-приложения, или так называемым сервером приложений.

Веб-приложение является связующим звеном между модулями комплекса. Его задачи – это авторизация и регистрация пользователей, генерация интерфейсных страниц (форм) для выполнения операций с комплексом, извлечение и сохранение необходимых данных из БД и БЗ. В связи с тем, что большинство рассмотренных операций имеют служебный характер и не являются вычислительно сложными, этот компонент реализован на динамическом языке Python с помощью фреймворка для создания веб-приложений Django [8]. Однако применение динамического языка за счет быстрого цикла отладки снижает время, затраченное на разработку. Соответственно, важным аспектом остается выбор эффективного веб-сервера. В настоящий момент это uWSGI как один из высокоэффективных серверов с поддержкой приложений на Python [9], при необходимости он легко может быть заменен другим без влияния на общую архитектуру.

Очередь задач – это компонент, отделяющий синхронную обработку задач, от асинхронных или потенциально ресурсоемких. Некоторые задачи, такие как предоставление приглашения для ввода пароля или отображение информации об одном ДСМ-эксперименте, выполняются быстро и за предсказуемое

время. Ожидание выполнения ДСМ-эксперимента (в зависимости от данных) может занять время экспоненциальное от объема входных данных. Для решения таких задач применяется технология отложенной обработки, в основе которой лежит очередь сообщений. Описание задач передается по сети в некотором оговоренном формате брокеру – программе, отвечающей за хранение и доступ к данным в очереди, а оттуда впоследствии извлекается обработчиками. Для текущей реализации брокером служит RabbitMQ [10] – популярная, хорошо масштабируемая и надежная система. Для обработчиков задач применяется система Celery для языка Python [11], что позволяет удобно соединять их с веб-приложением. Заметим, что это не нарушает принципов модульности и высокого быстродействия – основная работа выполняется решателем, а обработчики Celery выполняют лишь преобразование результата и подготовку данных. Еще одно преимущество использования очереди – это простое масштабирование комплекса для одновременного выполнения большего числа задач – множество программ обработчиков могут располагаться на различных машинах, извлекая задачи из общей очереди. Более того, создание прозрачно расширяемого кластера из программ обработчиков – типичное применение для системы Celery.

Важнейшим компонентом системы является решатель, непосредственно выполняющий процедуру генерации гипотез ДСМ-метода и предсказание неопределенных примеров по полученным гипотезам. В действительности эти две операции независимы и поэтому разделены на две различных программы – классификатор и генератор. Это дает больше гибко-

сти в применении: например, классификатор может провести предсказание примера, используя готовые гипотезы из базы знаний. Важная особенность данного модуля состоит в том, что необходимо соблюсти максимальную изолированность решателя от специфики предметной области, решатель самодостаточен и может быть использован в других проектах. С этой целью решатель и классификатор принимают и выдают данные об объектах в текстовом формате FIMI [12], не несущем дополнительной смысловой нагрузки (рис. 2). Формат FIMI – текстовый файл в кодировке ASCII: каждая строка описывает один объект и содержит разделенные пробелами номера атрибутов (или признаков), присутствующих у объекта, файл завершается пустой строкой. Единственной и достаточной внешней информацией является число атрибутов и признаков. Преобразование входных данных и результатов производится оболочкой, в нашем случае – это программы-обработчики Celery.

С точки зрения производительности ключевым свойством решателя является реализация алгоритмов параллельной генерации гипотез [13]. Параллельная обработка позволяет в несколько раз ускорить выполнение сложных экспериментов на больших выборках, эффективно используя доступные ресурсы многоядерных систем, типичных для современной серверной техники. В сочетании с масштабированием по числу одновременных задач на множестве узлов кластера, мы получаем достаточно широкие возможности для масштабирования, как для множества легких задач, так и для малого количества сложных. Помимо этих возможностей, чтобы обеспечить максимальную эффективность при ограниченных ресурсах, решатель и классификатор реализованы на языке C++11, с применением большинства известных техник для оптимизации аналогичных задач порождения [14-16].

Кодировщик – компонент, осуществляющий преобразование соединений из стандартного формата MDL (MOL файлы) в текстовые строки кода ФКСП. Для работы в составе комплекса кодировщик, представленный в [5], подвергся существенной доработке. По сравнению с исходной версией, появилась возможность получить не только коды, но и привязки – списков номеров атомов, соответствующих каждой найденной подструктуре. Привязки необходимы для предоставления эксперту-химику наглядной визуализации результатов работы комплекса, т. е. для перехода от кодов ФКСП обратно к (фрагментарной)

структурной формуле. Также для вывода этой структурированной информации используется стандартный формат данных – JSON (JavaScript Object Notation), это также радикально облегчает интеграцию с JavaScript в веб-браузере.

База пользователей отвечает за хранение учетных записей и сессий пользователей и используется веб-приложением. Без каких-либо существенных требований к этому компоненту была выбрана Sqlite как простая в эксплуатации реляционная СУБД [17]. Преимущество такого выбора – это отсутствие дополнительных программ-серверов и наличие поддержки в фреймворке Django, берущего на себя всю работу по управлению пользователями и сессиями, включая администрирование.

База фактов (БФ) хранит всю информацию, которую можно считать лишь данными, это сформированный эксперимент в виде списка соединений, свойств и активностей. База знаний же хранит результаты эксперимента в виде набора плюс- и минус-гипотез, предсказаний. Также в БЗ создан отдельный каталог по активностям и методу кодировки для гипотез, отобранных экспертами среди прочих. Критически важно соблюсти тот же метод кодирования соединения при использовании БЗ, поэтому каждое серьезное обновление кодировщика должно рассматриваться как новый метод кодирования. В нашей предметной области эксперимент и предсказания легко описываются иерархической моделью данных, поэтому, естественно, использовать СУБД, ориентированную на документы или объекты. Основными преимуществами такого подхода является возможность хранить сильно связанные данные в одном блоке.

Технически обе базы реализованы на основе NoSQL – документно-ориентированной СУБД MongoDB [18]. Одним из ее преимуществ является JSON-совместимого формата данных, что облегчает экспорт и импорт данных, взаимодействие с JavaScript. Также удобным для расширения системы является так называемая гибкая схема, при этом каждый документ определяет свою структуру сам, что дает возможность легко расширять любой из них дополнительной информацией. По мере роста объема данных встроенная в MongoDB поддержка кластеризации позволит эффективно распределить данные и нагрузку по нескольким серверам с незначительными изменениями в настройках веб-приложения.

```

1 3 9 13 23 25 34 36 38 40 52 54 59 63 67 76 85 86 90 93 98 107 113
2 3 9 14 23 26 34 36 39 40 52 55 59 63 67 76 85 86 90 93 99 108 114
2 4 9 15 23 27 34 36 39 41 52 55 59 63 67 76 85 86 90 93 99 108 115
1 3 10 15 23 25 34 36 38 41 52 54 59 63 67 76 85 86 90 93 98 107 113
2 3 9 16 24 28 34 37 39 40 53 54 59 63 67 76 85 86 90 94 99 109 114

```

Рис. 2. Формат данных FIMI, где строка описывает объект, а числа обозначают номера атрибутов и свойств, которыми он обладает

ПРИНЦИП РАБОТЫ СЕРВЕРА

Рассмотрим подробнее взаимодействие компонентов серверной части комплекса при выполнении основных операций: создание эксперимента, проведение эксперимента и классификация.

Особенность этапа создания эксперимента – загрузка пользователем файлов с химическим соединениями в интерактивном режиме, что достигается благодаря технологии AJAX (Asynchronous JavaScript and XML). Химические соединения – это достаточно сложные объекты, которые имеют множество стандартов представления. Речь идет как об обилии химических редакторов, различных форматов файлов и версий, так и о соглашениях. Примером соглашения является «неявный водород» – в таком случае водород в углеродных цепочках не записывается, но подразумевается. Open Babel – инструмент для конверсии химических файлов с открытым исходным кодом [19], который прекрасно решает такие стандартные задачи, как добавление «неявного» водорода и приведение файлов в единый формат. Полная схема обработки загружаемых пользователем соединений приведена на рис. 3. Заметим, что соединения хранятся в БД в пригодном для кодировки формате, и для каждого

уникального входного файла хранится ровно одна копия. Для исключения дубликатов в качестве ключа в БД используется 128-битная хеш-функция MD5 от содержимого исходного файла, тем самым также исключается шаг преобразования для одинаковых файлов.

Когда все молекулы добавлены и пользователь завершает ввод активностей, блок данных проверяется на корректность в веб-приложении. Затем все молекулы, указанные в виде идентификаторов (хеш-кодов), извлекаются из БД и передаются кодировщику (рис. 4) для обработки. Полученные коды и привязки добавляются к информации об активности и к файлу каждого исходного примера, а затем записываются в БД как единый документ.

Рассмотрим центральную задачу комплекса – процедуру проведения ДСМ-эксперимента. После запроса на проведение эксперимента, веб-приложение извлекает его данные и передает их в очередь, пользователь же перенаправляется на страницу ожидания результата (рис. 5). Обработка состоит из следующих фаз: разбиение входных примеров по типам, генерация (+)- и (-)-гипотез, предсказание неопределенных примеров из выборки, запись этих результатов в БД.

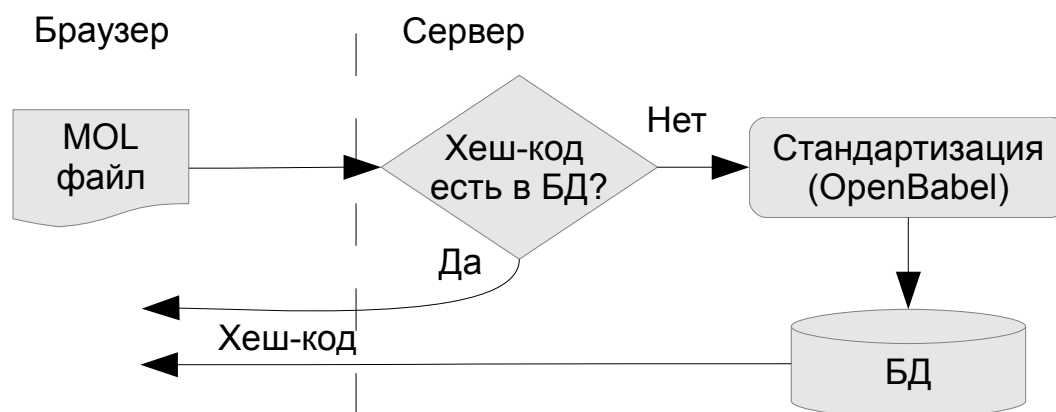


Рис. 3. Обработка загружаемых файлов с описанием молекул

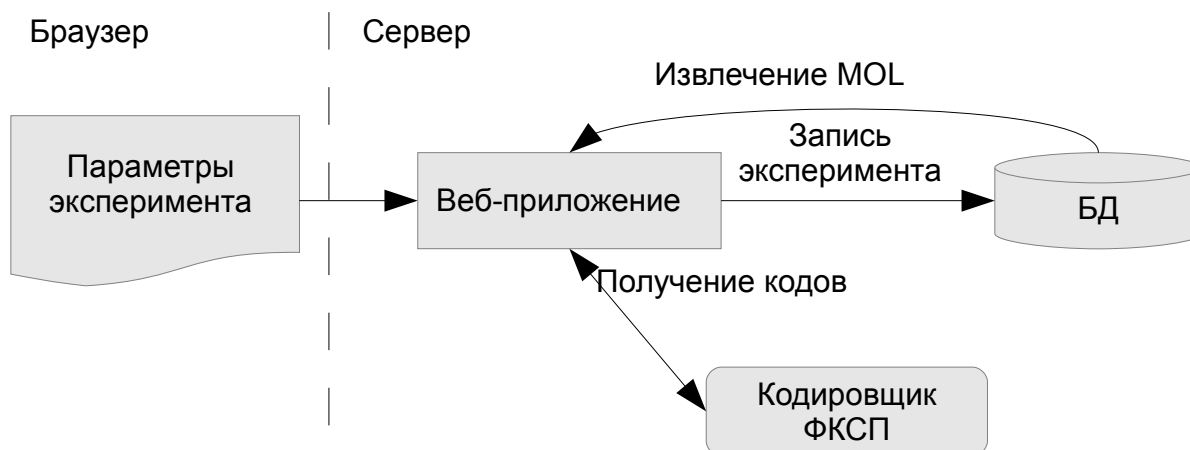


Рис. 4. Создание ДСМ-эксперимента

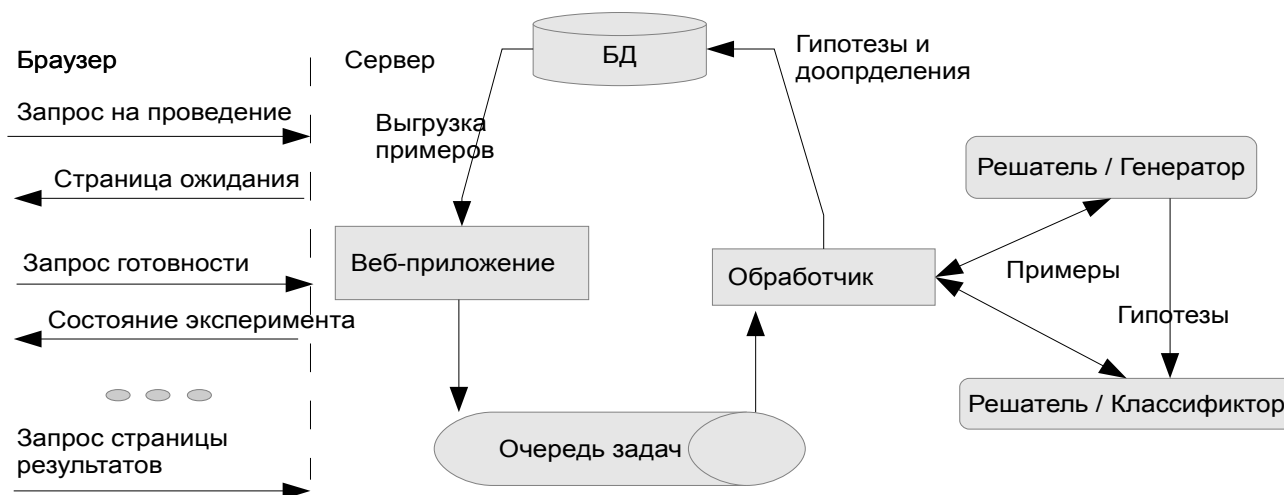


Рис.5. Проведение эксперимента

Первый шаг – это извлечение данных из очереди и разбиение примеров в зависимости от активности по трем группам: (+)-примеры, (-)-примеры и (t)-примеры. Активность в каждом эксперименте представлена как интервал $(0,1)$, разбитый на равные отрезки – степени, число степеней вводится при создании эксперимента. Активность любого примера представима строкой вида “++...+?...?--...-“, описывающей каждый отрезок от 0 до 1, слева направо. При ‘+’ – означает, что в этой степени активность проявляется, ‘-’ – не проявляется и ‘t’ – неизвестно. Очевидно, что если соединение проявляет активность степени A, то это верно и для всех степеней меньше A. Аналогично, если соединение не проявляет активность степени B, то и для всех последующих уровней – это верно. Отсюда и следует внешний вид нашего строчного описания – плюсу может предшествовать лишь плюс, за минусом – только минус, а неопределенность может быть лишь между ними. Разделение примеров происходит по следующей схеме: плюс-примеры – это примеры без t-компоненты и хотя бы с одним плюсом (без ‘?’), для которых минусы заменяются нулями. Для минус-примеров аналогично, но плюсы замещаются нулями и должен быть хотя бы один минус. Иными словами, любой пример без неопределенностей с активностью $activity > 0$ – положительный пример для всех степеней ниже $activity$. С активностью $activity < 1$ – отрицательный, для всех степеней выше $activity$. И, наконец, в t-примеры попадает любой пример с t-компонентой, без изменений. Все три группы примеров записываются в формате FIMI по трем различным файлам.

Генерация гипотез опирается на разработанный ранее параллельный решатель [13], который запускается одновременно для двух наборов примеров – плюс и минус. В качестве параметра передается стратегия ДСМ-метода – прямой алгоритм или с запретом на контр-примеры. В первом алгоритме из рассмотрения исключаются гипотезы, для которых есть пример с такими же атрибутами и свойствами строго противоположного знака. Запрет на контр-примеры расширяет это правило – гипотезы исключаются, если найдется пример, в который вкладываются атрибуты гипотезы, и имеющий свойства строго противо-

положного знака. Полученные гипотезы передаются вместе с t-примерами классификатору. Для каждого примера выдается список вкладывающихся в него гипотез и выведенную строку активности, включая противоречия, которые записываются как нули.

Завершающим шагом обработчик записывает в формате JSON гипотезы и предсказания в БЗ для данного эксперимента.

ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС

Перейдем к внешнему представлению – веб-страницам, предоставляющим интерфейс пользователю-эксперту. Основные страницы – просмотр списка экспериментов, создание эксперимента, просмотр данных одного эксперимента, просмотр результатов: гипотез и доопределений, анализ по БЗ. Для наглядности, наиболее значимые из них показаны на рис. 6-9.

Главная страница – список экспериментов (рис. 6). Она служит каталогом всех экспериментов, проводимых пользователем ранее, и отправной точкой для дальнейшей работы. Основные действия – просмотр краткой сводки по любому из экспериментов, создание нового эксперимента, переход к деталям конкретного эксперимента или его удаление. Есть также возможность создать новый эксперимент на основе текущего, при этом входные данные текущего будут уже внесены в форму.

Создание эксперимента (рис. 7) представлено краткой формой для ввода данных – название, активность, количество уровней активности и дополнительно комментария. Центральное место занимают кнопки загрузки файлов с примерами, а также таблица для редактирования уже загруженных соединений. Благодаря технологии AJAX для пользователя это выглядит как интерактивный режим традиционных графических приложений. Редактирование позволяет вручную задать строку активности или удалить лишний пример из списка. Альтернативой ручному вводу с помощью кнопок является импорт активностей из файла с описанием эксперимента прежней системы [20], либо стандартного файла таблиц в формате CSV. Если пренебречь тонкостями формата, то все такие файлы содержат строки вида: имя файла – строка активности, которые сопоставляются списку загруженных соединений.

Список экспериментов
Анализировать по БЗ
admin

Просмотр экспериментов

Список экспериментов

- Галогены
- Галогены-2
- Ливерпуль

Новый +

Название: Галогены-2
Дата: 11.06.2015

Активность: Мутагенность
Уровень активности: 1

Метод: Прямой
Кодировщик: fcsp-v1

Соединений: 63
Гипотез "+": 0
Гипотез "-": 0

Комментарий

Перейти к эксперименту

Новый на основе текущего

Удалить эксперимент

Рис.6. Страница просмотра экспериментов

Список экспериментов
Анализировать по БЗ
admin

Новый эксперимент

Название

Дата

Активность

Уровень активности

Метод

Кодировщик

Комментарий

Загрузить MOL файлы

Импортировать активность

Создать эксперимент

Название	Активность	
10.MOL	+	+ -
11.MOL	-	+ -
12.MOL	+	+ -
13.MOL	?	+ -

Рис.7. Страница создания эксперимента

Эксперимент "Ливерпуль"

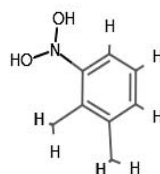
Данный эксперимент был выполнен. Перейти к гипотезам.

Входные данные

Выберите пример из списка, чтобы получить детальную информацию.

2	100.MOL	??
3	101.MOL	--
4	102.MOL	--
5	103.MOL	--
6	104.MOL	--
7	105.MOL	--
8	106.MOL	--
9	107.MOL	--
10	108.MOL	--

Схема молекула и коды



6,06 0300150 0300150 0300331 3301411
3301411 4162411 0363410

Рис. 8. Просмотр данных эксперимента

Гипотезы

Гипотезы		Коды	Родители	Молекулы
1 (-)	+БЗ	3303411	30.MOL(178)	Родитель
2 (-)	+БЗ	0300331	81.MOL(234)	
3 (-)	+БЗ	0300150		
4 (-)	+БЗ	0300150		
5 (-)	+БЗ	6,06		
6 (-)	+БЗ			

Рис.9. Страница просмотра гипотез и доопределений

После создания эксперимента или при просмотре деталей существующего, пользователь попадает на страницу просмотра эксперимента. В этой странице отображается набор входных данных эксперимента, коды ФКСП и визуализация выбранного примера (рис. 8). Таким образом, можно изучить выборку перед отправкой ее решателю, чтобы оценить успешность кодирования всех молекул из набора. Отсутствие кодов или их низкая плотность может указать на потребность расширить код ФКСП, добавив новые дескрипторные центры [1, 5], соответственно зарегистрировав новые параметры в системе как еще один способ кодирования.

Из страницы просмотра переходим к странице просмотра гипотез и доопределений. Если эксперимент еще не был выполнен, то пользователю перед просмотром придется некоторое время подождать на странице ожидания. На практике проведение большинства старых экспериментов на новом комплексе занимает менее секунды. Страница просмотра делится на три вкладки: плюс-гипотезы, минус-гипотезы и предсказания, а также предоставляет богатые возможности для визуальной оценки гипотез экспертом. Вкладки предсказаний содержат список гипотез. Для каждой выбранной гипотезы отображается перечень кодов ФКСП и список родителей, также визуализируются выбранная гипотеза и выбранный родитель (рис. 9). Вкладка t-примеров выглядит аналогично, но вводит еще один уровень выбора – для каждого из примеров приводится список гипотез, относящихся к нему, и уже для этого списка все выглядит аналогично вкладкам гипотез.

Со страницы эксперимента добавление гипотез в БЗ осуществляется кликом по знаку “БЗ+”, удаление – по “БЗ-”. Страница анализа соединений по БЗ представляет собой расширение страницы просмотра гипотез – та же информация приведена для набора загруженных для анализа соединений, но предоставляется несколько страниц по каждой активности и гипотезы для классификации извлекаются из БЗ.

Общей подзадачей многих форм пользовательского интерфейса является визуализация гипотезы. Как уже упоминалось, гипотеза в нашем случае — это набор слов, который описывает отдельные подструктуры, при этом информация о расположении этих подструктур в каждом из родителей гипотезы становится неоднозначной. Даже при наличии привязок кодов к родителю, мы не можем однозначно определить точный вид гипотезы. Так при кратности кода в гипотезе, равной двум, для родителя с четырьмя вхождениями данного кода имеется $C_4^2 = 6$ вариантов расположения этих подструктур в родителе.

Таким образом, чтобы дать эксперту максимально полную картину, визуализация выполняется в виде ряда изображений, учитывающих все возможные расположения фрагментов в каждом из родителей гипотезы по очереди. Наиболее экономичным решением с точки зрения объема передаваемых данных и сокращения лишних вычислений оказалось отсеечение лишних атомов из родителей прямо в браузере пользователя. Генерация изображения производится

только в момент выбора очередной комбинации, достаточная для этого информация – это группы номеров атомов и родительская молекула.

ЗАКЛЮЧЕНИЕ

Созданный программный комплекс является инструментом для проведения исследований в задаче выявления связи структура-активность, для фармакологии и смежных областях. Сочетание БЗ и БД вместе с удобным пользовательским интерфейсом, механизмами распределения нагрузки, кодировщиком и решателем представляет собой платформу для дальнейшей работы специалистов в области искусственного интеллекта и химии. Также комплекс реализует инструмент для фильтрации гипотез в БЗ, классификации на основе знаний и является инструментом для интеллектуального анализа данных (ИАД).

Модульная архитектура комплекса позволяет легко интегрировать другие версии компонентов – ДСМ-решателей и кодировщиков химических соединений. Таким образом, создается среда для сравнительного анализа различных реализаций ДСМ. Основные решения, принятые в построении комплекса, способствуют высокой эффективности, гибкости, простоте настройки, высокой доступности и централизации информации.

В настоящий момент на приведенном программном комплексе повторяются эксперименты, проведенные ранее. После анализа результатов этих испытаний и выявления возможных дефектов, комплекс станет новым инструментом для проведения исследований структура-активности в ВИНТИ РАН. Доступ для испытаний сторонним исследователям предоставляется по запросу.

СПИСОК ЛИТЕРАТУРЫ

1. Cunningham A.R., Cunningham S.L., Consoer D.M., Moss S.T., Karol M.H. Development of an information-intensive structure-activity relationship model and its application to human respiratory chemical sensitizers // SAR and QSAR in Environmental Research. – 2005. – Vol. 16, № 3. – P. 73-285
2. Kumar P., Carrasquer C.A., Carter A., Song Z., Cunningham A.R. A categorical structure-activity relationship analysis of GPR119 ligands // SAR and QSAR in Environmental Research. – 2014. – Vol. 25, № 11. – P. 891-903.
3. Лейбов А.Е. Автоматическое кодирование химических структур кодом ФКСП. // Итоги науки и техники. Сер. Информатика. – 1991. – Т. 15. – С.141-158.
4. Блинова В.Г., Добрынин Д.А. Языки представления химических структур интеллектуальных системах для конструирования лекарств // Научно-техническая информация. Сер. 2. – 2000. – № 6. – С. 14-21.
5. Ольшанский Д.Л., Блинова В.Г., Добрынин Д.А. Автоматическое кодирование химиче-

- ских соединений фрагментарным кодом суперпозиции подструктур // Научно-техническая информация. Сер.2. – 2015. – № 5. – С. 7-18.
6. Sushko I., Salmina E., Potemkin V.A., Pod G., Tetko I. V. ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds with potential adverse reactions // Journal of chemical information and modeling. – 2012. – Vol. 52.
 7. HTML5 W3C Recommendation. – URL: <http://www.w3.org/TR/html5/>
 8. The Django project. – URL: <https://www.djangoproject.com/>
 9. The uWSGI project. – URL: <http://uwsgi-docs.readthedocs.org/en/latest/>
 10. RabbitMQ. – URL: <https://www.rabbitmq.com/>
 11. Celery: Distributed Task Queue. – URL: <http://www.celeryproject.org/>
 12. Frequent Itemset Mining Dataset Repository. – URL: <http://fimi.ua.ac.be/data/>
 13. Ольшанский Д.Л. Подбор алгоритма для параллельной реализации метода сходства в интеллектуальных ДСМ-системах // Научно-техническая информация. Сер.2. – 2015. – №7. – С. 10-19.
 14. Krajca P., Outrata J., Vychodil V. Parallel Recursive Algorithm for FCA. // Proc. CLA CEUR WS, 433 — 2008
 15. Krajca P., Outrata J., Vychodil V. Advances in Algorithms Based on CbO // Proc. CLA. – 2010. – P. 325-337.
 16. Andrews S. A 'Best-of-Breed' approach for designing a fast algorithm for computing fixpoints of Galois Connections // Information Sciences. – 2015. – P. 633-649.
 17. Sqlite. – URL: <https://www.sqlite.org/>
 18. MongoDB. – URL: <https://www.mongodb.org/>
 19. Open Babel: The Open Source Chemistry Toolbox. – URL: <http://openbabel.org/>
 20. Блинова В.Г., Добрынин Д.А, Панкратова Е.С, Фабрикантова Е.Ф., Финн В.К. Интеллектуальные системы типа ДСМ – компьютерная реализация системного анализа в науках о жизни // Труды Первой международной конференции САИТ-2005. Переславль-Залесский, 12-16 сентября 2005 г., Т.1. – М: Ком- книга, 2005. – С. 149-155.

Материал поступил в редакцию 16.06.15.

Сведения об авторах

ОЛЬШАНСКИЙ Дмитрий Леонидович – аспирант Всероссийского института научной и технической информации Российской академии наук, Москва
e-mail: dmitry.olsh@gmail.com

В «Литературной газете» (№ 31 от 29 июля – 4 августа 2015 г.) опубликована статья французского коллекционера Р. Герра, содержащая ложные сведения относительно событий тридцатилетней давности в Библиотеке Российской академии наук, оскорбительные для ее директора В.П. Леонова и всего информационно-библиотечного сообщества.

Редакция и редакционные коллегии сборника «Научно-техническая информация» выражают крайнее неодобрение газете и полную поддержку нашему коллеге доктору педагогических наук, профессору Валерию Павловичу Леонову, которого знаем и уважаем много лет как серьезного ученого, успешного руководителя, интеллигентного и высокопорядочного человека.

Главный редактор Р.С. Гиляревский

О корректности ДСМ-рассуждений в медицинских системах*

В Отделе теоретических и прикладных проблем информатики ВИНТИ РАН были созданы интеллектуальные ДСМ-системы в области медицины – «Прогнозирование послеоперационных осложнений» и «Прогнозирование развития осложнений хронического панкреатита». Дается описание исследования этих систем с точки зрения корректности проводимых в них рассуждений.

Ключевые слова: интеллектуальная система, ДСМ-метод, медицинская диагностика

ДСМ-системы для прогнозирования послеоперационных осложнений [1] и развития осложнений хронического панкреатита [2] используют разные виды признаков и подбор количества «родителей»¹ для нахождения гипотез первого рода (гипотез о причинах явлений). Подбор этого порога осуществляется по принципу наилучшего результата в смысле полноты и точности прогноза при применении процедуры «доопределение по одному» [1].

Проведенные эксперименты позволили выявить некоторые эмпирические зависимости развития осложнений от наличия у больного конкретного набора признаков и подтвердили возможность использования этих ДСМ-систем для медицинской диагностики.

Однако сам факт нахождения эмпирической зависимости еще недостаточен, чтобы сделать вывод о том, что с помощью ДСМ-системы найдена эмпирическая закономерность или выявлена некоторая тенденция. Поскольку ДСМ-метод использует правдоподобные рассуждения, то и выводы, получаемые в результате проведения этих рассуждений, являются правдоподобными. ДСМ-системы – это открытые системы, они допускают расширение баз фактов (БФ). Более того, если не выполняется критерий достаточного основания [3], то идеология ДСМ-метода требует расширения базы фактов специально подобранными примерами.

Пусть при работе с базой фактов БФ₁ было получено множество гипотез о причинах явлений (гипотезы I рода):

$\Delta_1 = (\Delta_1^+ \cup \Delta_1^- \cup \Delta_1^0)$, где Δ_1^+ – положительные, Δ_1^- – отрицательные, Δ_1^0 – противоречивые гипотезы I рода,

и множество гипотез о доопределении примеров (гипотезы II рода):

$\Omega_1 = (\Omega_1^+ \cup \Omega_1^- \cup \Omega_1^0)$, где Ω_1^+ – положительные, Ω_1^- – отрицательные, Ω_1^0 – противоречивые гипотезы II рода.

В зависимости от применяемых стратегий при расширении БФ₁ до БФ₂ возможны следующие ситуации, связанные с соотношением множеств гипотез Δ_1 и Ω_1 , полученных из БФ₁, и Δ_2 и Ω_2 , полученных из БФ₂:

1. $(\Delta_1^+ \cap (\Delta_2^- \cup \Delta_2^0)) = \emptyset$ & $(\Delta_1^- \cap (\Delta_2^+ \cup \Delta_2^0)) = \emptyset$ & $(\Delta_1^0 \cap (\Delta_2^+ \cup \Delta_2^-)) = \emptyset$ и $(\Omega_1^+ \cap (\Omega_2^- \cup \Omega_2^0)) = \emptyset$ & $(\Omega_1^- \cap (\Omega_2^+ \cup \Omega_2^0)) = \emptyset$ & $(\Omega_1^0 \cap (\Omega_2^+ \cup \Omega_2^-)) = \emptyset$.

2. $(\Delta_1^+ \cap \Delta_2^0) = \emptyset$, но $(\Omega_1^+ \cap \Omega_2^0) \neq \emptyset$, где $\sigma \in \{+, -, 0\}$, $\rho \in \{+, -, 0\}$, σ и ρ – различны.

3. $(\Delta_1^+ \cap \Delta_2^0) \neq \emptyset$ и $(\Omega_1^+ \cap \Omega_2^0) \neq \emptyset$, где $\sigma \in \{+, -, 0\}$, $\rho \in \{+, -, 0\}$, σ и ρ – различны.

Вариант $(\Omega_1^+ \cap \Omega_2^0) = \emptyset$ и $(\Delta_1^+ \cap \Delta_2^0) \neq \emptyset$ невозможен, так как из $(\Omega_1^+ \cap \Omega_2^0) = \emptyset$ следует $(\Delta_1^+ \cap \Delta_2^0) = \emptyset$. (Доказательство этого факта см. в [4]).

Ситуация 1 означает, что при расширении баз фактов не возникает противоречий между гипотезами, полученными раньше, и новыми. Из этого следует корректность ДСМ-рассуждений для данной стратегии [5].

Ситуация 2 означает корректность I рода [6], т.е. непротиворечивость гипотез при расширении баз фактов только для гипотез I рода.

Если имеет место Ситуация 3, то проводимое в данной стратегии ДСМ-рассуждение – некорректно. Впрочем, в этом случае важно количество противоречий. Если их меньше определенного порога, то можно говорить о выявлении тенденции [7].

В связи со сказанным, рассуждения, проводившиеся в ДСМ-системах для прогнозирования послеоперационных осложнений и развития осложнений хронического панкреатита, были проверены на корректность.

Прежде чем изложить полученные в ходе этой проверки результаты, опишем кратко сами указанные ДСМ-системы.

В ДСМ-системе «Прогнозирование послеоперационных осложнений» решалась задача: по данным об онкологических пациентах, содержащимся в исто-

* Работа выполнена при поддержке программы фундаментальных исследований Президиума РАН (П15, проект №209) и гранта РФФИ (проект 14-07-00856а).

¹ «Родители» – объекты из базы фактов, порождающие гипотезу.

рии болезни, включающим анамнез и результаты клинического исследования образцов крови, предсказать, будет ли после операции хотя бы одно осложнение из указанного списка.

В БФ содержатся сведения о больном, включающие данные лабораторных исследований до и после операции, сопутствующие заболевания и т.д.

ДСМ-система «Прогнозирование развития осложнений хронического панкреатита» была создана для прогнозирования сахарного диабета панкреатогенного генеза. База фактов создана на основе информации из историй болезни пациентов. Каждый объект в БФ представлен в виде кортежа из 39 элементов.

Для проведения ДСМ-рассуждений в обеих системах были выбраны две однородные стратегии – стратегия с (+ и -)-предикатами простого метода сходства и стратегия с (+ и -)-предикатами простого сходства с запретом на контрпримеры.

Чтобы уточнить порог для минимального количества «родителей», необходимого для получения гипотез, применялась процедура «доопределение по одному», которая состоит в том, что каждому объекту из БФ последовательно присваивается значение «неизвестно» и этот объект доопределяется. Результат сравнивается с реальным фактом из БФ. В результате проведения этой процедуры для всех объектов из БФ при разном количестве «родителей», выбирается такой порог количества «родителей», который дает наилучшую точность и приемлемую полноту. Далее реализуется ДСМ-метод, причем кандидатами в гипотезы считаются только сходства объектов в количестве не менее выбранного порога «родителей».

Для проверки корректности проводимых ДСМ-рассуждений в каждой системе были выбраны три расширяющиеся базы фактов: БФ₁, БФ₂ и БФ₃.

Для перехода от БФ₁ к БФ₂ и от БФ₂ к БФ₃ нужно завершить соответствующий этап ДСМ-рассуждения, т.е. повторить такты (индукция → аналогия) до стабилизации [7] и провести абдуктивное рассуждение, чтобы убедиться, что найденные гипотезы объясняют базу фактов. В рассматриваемой системе стабилизация наступает уже на первом такте, т.е. множество гипотез, порожденных на такте 1, совпадает с множеством гипотез, порожденных на такте 2.

Основная сложность при работе с данной системой состоит в подборе порога количества «родителей». Дело в том, что с ростом мощности базы фактов сильно растет количество гипотез I рода, поэтому порог количества «родителей» растет вместе с ростом мощности БФ. При этом порог, подходящий для самой большой БФ, может дать пустое множество гипотез для самой маленькой БФ. Таким образом приходится тщательно подбирать порог количества «родителей», чтобы получить хорошую точность и полноту при «доопределении по одному» и приемлемое количество гипотез в расширяющихся базах фактов.

Для системы «Прогнозирование послеоперационных осложнений» были использованы базы фактов с мощностями соответственно: |БФ₁| = 22, |БФ₂| = 40, |БФ₃| = 49.

Гипотезы, найденные в БФ₁, БФ₂ и БФ₃ с применением однородной стратегии с использованием простого метода сходства, не изменяли знака при пере-

ходе к расширенной базе фактов. Это значит, что для ДСМ-рассуждения, проводимого в данной ДСМ-системе с использованием простого метода сходства, выполняется корректность I рода.

Проверка ДСМ-рассуждений, проводимых с помощью стратегии с использованием простого метода сходства с запретом на контрпримеры, на корректность I рода смысла не имеет, так как метод с запретом на контрпримеры не может давать противоречивые результаты для гипотез I рода и, следовательно, корректность I рода всегда выполнена в этом случае. Однако при нахождении гипотез с помощью однородной стратегии с использованием метода сходства с запретом на контрпримеры в расширяющихся базах данных, восемь положительных гипотез, полученных из БФ₂ при переходе к БФ₃, не были найдены. Это значит, что в БФ₃ нашлись контрпримеры, содержащие эти гипотезы.

Эта ситуация свидетельствует о том, что, хотя корректность I рода для рассматриваемой стратегии в силу ее определения всегда имеет место, но более сильное условие – сохранение истинности гипотез при расширении базы фактов, в данном случае места не имеет.

Сохранение истинности гипотез в ДСМ-рассуждении [8] имеет место тогда, когда выполнено: $(\Delta^{\sigma}_i \cap \Delta^{\rho}_j) = \emptyset \ \& \ (\Omega^{\sigma}_i \cap \Omega^{\rho}_j) = \emptyset$, где $\sigma \in \{+, -, 0, \tau\}$, $\rho \in \{+, -, 0, \tau\}$, σ и ρ – различны. Очевидно, что для простого метода сходства из корректности ДСМ-рассуждения следует и сохранение истинности. Но для простого метода с запретом на контрпримеры это не так. Если при расширении базы фактов появляется контрпример для положительной или отрицательной гипотезы, то истинностное значение этой гипотезы становится равным τ (неопределенность), т.е. фактически эта гипотеза на новом этапе исчезает из множества гипотез, что и произошло с восемью положительными гипотезами в ДСМ-системе «Прогнозирование послеоперационных осложнений».

Поскольку порог количества «родителей» в данном случае для отрицательных примеров равен 10, то не найденные положительные гипотезы теоретически могли содержаться в БФ₃ максимум в девяти объектах из отрицательных примеров. Но просмотр БФ₃ показал, что все восемь исчезнувших при расширении базы фактов гипотез содержатся в одном и том же объекте из отрицательного примера максимальной базы фактов. (Этот пример появляется только в БФ₃, поэтому при переходе от БФ₁ к БФ₂ все гипотезы сохранили истинностное значение). Дальнейший анализ данных показал, что объект, являющийся контрпримером, содержит признак под номером 13 со значением, встречающимся во всех положительных примерах и не встречающимся ни в одном отрицательном примере, кроме того, который оказался контрпримером. Это наводит на мысль, что либо значение признака №13 в контрпримере заполнено ошибочно, либо этот пример ошибочно означен как отрицательный. В силу технических сложностей выяснения реального положения дел в данном случае пришлось попытаться понять это косвенным путем. Для этого была выполнена процедура «доопределение по-одному» в двух вариантах. Первый вариант –

значение признака 13 в контрпримере было исправлено в соответствии с тем, каково оно в отрицательных примерах (этот признак имеет два значения). Второй вариант – контрпример был переведен из множества отрицательных во множество положительных примеров. Результаты сравнивались с результатами данной процедуры при исходном варианте. Оказалось, что изменение значения признака в контрпримере дало результаты «доопределения по одному» лучше, чем они были в исходной базе фактов, а изменение знака примера ухудшило исходные доопределения. Из этого можно сделать вывод, что ошибка, скорее всего, в значении признака. Таким образом, в данной системе стратегии с простым методом сходства и простым методом сходства с запретом на контрпримеры дают одинаковые результаты. Кроме того, выявляется и еще один важный факт – признак №13 является разделяющим признаком для выявления послеоперационных осложнений.

Для системы «Прогнозирование развития осложнений хронического панкреатита» были использованы базы фактов с мощностями, соответственно: $|БФ_1| = 81$, $|БФ_2| = 123$, $|БФ_3| = 153$.

При проведении ДСМ-рассуждений в этой системе с расширяющимися базами фактов ситуация складывалась такая же, как и в системе «Прогнозирование послеоперационных осложнений». Имеет место корректность I рода. При использовании стратегий с простым методом сходства и с простым методом сходства с запретом на контрпримеры, результаты совпали.

Проведенные эксперименты и их анализ позволили сделать вывод о том, что ДСМ-рассуждения, проводимые в интеллектуальных системах «Прогнозирование послеоперационных осложнений» и «Прогнозирование развития осложнений хронического панкреатита» с помощью стратегий с простым методом сходства и с простым методом сходства с запретом на контрпримеры, обладают корректностью I рода, сохраняют истинные значения гипотез и дают в каждой системе одинаковые результаты для этих стратегий.

Анализ результатов позволил выявить ошибку в базе фактов и обнаружить разделяющий положительные и отрицательные примеры признак.

СПИСОК ЛИТЕРАТУРЫ

1. Панкратова Е. С., Добрынин Д. А. Компьютерная интеллектуальная ДСМ-система для прогнозирования послеоперационных осложнений // Научно-техническая информация. Сер.2. – 2015. – № 6. – С.9-13.
2. Винокурова Л.В., Агафонов М.А., Варварина Г.Г., Финн В.К., Панкратова Е.С., Добрынин Д.А. Применение интеллектуальной системы типа ДСМ для анализа клинических данных // Российский Биотерапевтический Журнал. – 2014. – Т. 13, №3. – С.57-60.
3. Финн В. К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. Сер. «Информатика». – 1991. – Т. 15.
4. Гусакова С. М. Корректность ДСМ-рассуждений для однородных стратегий // Научно-техническая информация. Сер.2. – 2015. – № 5. – С. 1-6.
5. Финн В. К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть II // Искусственный интеллект и принятие решений. – 2010. – №4. – С.14-40.
6. Гусакова С. М. Отношение толерантности, характеризующее корректность ДСМ-рассуждений // Труды XIII национальной конференции по искусственному интеллекту с международным участием – КИИ-2012 (16-20 октября 2012 г.). – Белгород, 2012. – С.100–107.
7. Финн В. К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть I // Искусственный интеллект и принятие решений. – 2010. – №3. – С.3-21.
8. Финн В. К. Обнаружение эмпирических закономерностей посредством ДСМ-рассуждений // Научно-техническая информация. Сер.2. – 2015. – № 8. – С. 1-29.

Материал поступил в редакцию 30.06.15.

Сведения об авторах

ГУСАКОВА Светлана Марковна – кандидат физико-математических наук, старший научный сотрудник сектора интеллектуальных информационных систем ВИНТИ РАН, Москва
e-mail: smg@viniti.ru

ПАНКРАТОВА Елена Сергеевна – кандидат технических наук, старший научный сотрудник сектора интеллектуальных информационных систем ВИНТИ РАН.
E-mail: pankr@viniti.ru

М.Р. Когаловский, С.И. Паринов

Таксономия семантических связей информационных объектов контента научной электронной библиотеки*

Обсуждается таксономия семантических связей, позволяющая авторизованным пользователям научной электронной библиотеки в онлайн-режиме декларировать различные виды отношений между информационными объектами ее контента. В статье рассматривается организация таксономии и ее представление в виде набора контролируемых словарей классов связей. Описывается реализация таксономии в среде крупной научной информационной системы Соционет, а также ряд новых возможностей для сообщества пользователей системы, обеспечиваемых благодаря семантическому обогащению ее контента путем спецификации семантических связей.

Ключевые слова: научная электронная библиотека, контент библиотеки, система Соционет, семантическая связь, таксономия связей, контролируемый словарь, новые функции электронной библиотеки

ВВЕДЕНИЕ

В последние годы в области научных электронных библиотек активно развиваются исследования и разработки, направленные на семантическое обогащение их контента. Благодаря этому может быть повышена информативность, а, следовательно, и ценность информационных ресурсов, представленных в этих системах. Помимо традиционного доступа пользователей непосредственно к информационным объектам библиотеки, открываются новые возможности для практического применения семантически обогащаемых электронных библиотек и решения ряда новых нетрадиционных задач.

Один из активно развивающихся подходов к семантическому обогащению информационных ресурсов научных электронных библиотек, на котором основаны и работы авторов данной статьи, заключается в предоставлении пользователям возможностей декларации *семантических связей* между информационными объектами контента библиотеки и визуализации созданных связей. Такие связи могут нести различную информацию, определенную явным образом. Они могут представлять разнообразные отношения научного характера между публикациями, в частности, констатировать факт цитирования с указанием его мотивов, характеризовать влияние одной публикации на другую или использование в ней результатов, обсуждаемых в другой публикации, определять структурные связи между компонентами

или различными версиями научного произведения, представлять отношения авторства между персонами и публикациями, декларировать вклад соавторов в создание коллективной публикации и др. Могут представляться также отношения между пользователями системы (представленными их персональными профилями) и публикациями библиотеки. Например, пользователи библиотеки могут декларировать оценки представленных в ней публикаций, выражать рекомендации их авторам или публиковать свое мнение относительно информации, содержащейся в созданных семантических связях.

Семантические связи, кроме связей цитирования и некоторых других, в предлагаемом подходе декларируются *децентрализованно авторизованными пользователями* электронной библиотеки в онлайн-режиме, имеющими при этом дело со своеобразной социальной сетью. Семантика связей цитирования может декларироваться только пользователями-авторами цитирующих публикаций. Возможности пользователей декларировать новые семантические связи в онлайн-режиме в процессе их работы с контентом библиотеки приводят к формированию динамически обогащаемого «*семантического ореола*» библиотеки, его *многослойной семантической структуры*, слои которой соответствуют различным классам декларированных семантических связей [1–3].

Научная электронная библиотека с семантически структурированным контентом может служить основой для поддержки новых видов научных коммуникаций [4], информационным источником для качественно новых наукометрических измерений и иссле-

* Работа поддержана грантом РГНФ 14-02-12010-в.

дований на представленном в библиотеке корпусе научных знаний, а также для ряда других целей. Эти возможности и разработанные механизмы для их практического применения подробно обсуждаются в работах [1, 3, 5, 6–10].

Описываемая в настоящей статье таксономия семантических связей является одним из ключевых элементов предложенного и реализованного авторами подхода к созданию семантически обогащаемых научных электронных библиотек.

В принятом авторами подходе допускаются только бинарные ориентированные семантические связи между ее информационными объектами. Объект – участник связи, из которого исходит связь, называется ее *исходным* информационным объектом. Второй участник, на который направлена связь, называется *целевым* объектом этой связи.

Семантические связи в предлагаемом подходе представляются в электронной библиотеке как *самостоятельные информационные объекты* [1, 5]. Автономное представление семантических связей позволяет использовать объекты-связи как участников других связей. Кроме того, благодаря этому можно формировать автономные репозитории семантических связей, которые могут реализовываться как самостоятельные источники информации и могут интегрироваться как с другими подобными репозиториями, так и с различными электронными библиотеками, если они используют единый с данной библиотекой стандарт представления метаданных и общий с ней метод идентификации информационных объектов.

В качестве полигона для реализации рассматриваемого подхода используется крупная отечественная *научная информационная система Соционет* [11], функционирующая уже полтора десятилетия. Соционет основана на технологии открытых архивов [12, 13], импортирует метаданные из репозитория более 1700 организаций, включая сотни отечественных и зарубежных открытых архивов, предоставляет метаданные собственного репозитория для импорта другим системам и т.д. В настоящее время в Соционет доступно около 2,5 млн информационных объектов, содержатся описания более 8 млн семантических связей.

В остальной части статьи обсуждаются известные средства определения семантики декларируемых связей (разд. 1), на которые опирается предлагаемый подход, дается общая характеристика предлагаемой таксономии (разд. 2) и ее представления в виде контролируемых словарей (разд. 3), описывается техника создания семантических связей в среде системы Соционет (разд. 4). Наконец, кратко обсуждаются возможности, которые обеспечиваются в системе для пользователей благодаря поддержке семантической структуры связей и принятому способу их представления в системе (разд. 5).

1. СРЕДСТВА ОПРЕДЕЛЕНИЯ СЕМАНТИКИ ДЕКЛАРИРУЕМЫХ СВЯЗЕЙ

Декларация семантики связей в созданной авторами технологии осуществляется пользователями электронной библиотеки на основе *таксономии связей*. При разработке таксономии были приняты во внимание и использованы результаты ряда известных

проектов последних лет. Наиболее продвинутые работы в этой области были выполнены специалистами в области биомедицины.

Одну из ранних фундаментальных разработок представляет собой модульный комплекс онтологий *SPAR (The Semantic Publishing and Referencing Ontologies)* [14, 15], созданный сотрудниками Оксфордского и Болонского университетов. Этот комплекс включает восемь независимых онтологий, позволяющих описывать семантику библиографических объектов, а также их отношений, в частности библиографических объектов, библиографических записей и источников в пристатейных списках, связи цитирования, контексты цитирования и их связи с релевантными разделами цитируемых публикаций. Кроме того, онтологии комплекса SPAR могут использоваться для описания семантики компонентов документов, ролей и состояний публикаций, потоков работ в издательских процессах.

Другой комплекс онтологий – *SWAN (Semantic Web Applications in Neuromedicine)* [16] создан специалистами в области нейромедицины в Главном госпитале Массачусетса и Медицинской школе Гарварда. Комплекс SWAN предназначен для обеспечения комфортной среды в Семантическом Вебе, названной авторами *социально-технической экосистемой*. Эта среда позволяет создавать и сохранять семантический контекст научных коммуникаций, обеспечивает доступ к нему, его интеграцию, а также обмен неструктурированной и слабоструктурированной цифровой научной информацией. В онтологиях SWAN классы связей более агрегированы по сравнению с классами SPAR.

В данном контексте заслуживают внимания разработанные консорциумом W3C рекомендации *SKOS (Simple Knowledge Organization System)* [17]. Разработчики SKOS создавали ее для использования в системах организации знаний – тезаурусах, схемах классификации, таксономиях и рубрикаторах (*Subject Heading Systems*), поддерживаемых в среде Семантического Веба. Рекомендация SKOS определяет концептуальную схему, называемую *общей моделью данных*, которая позволяет совместно использовать и связывать различные системы организации знаний в среде Веба. При использовании такой унифицированной концептуальной схемы упрощается интеграция существующих систем организации знаний в среде Семантического Веба.

В научных электронных библиотеках значительна доля коллективных работ. Поэтому представляет интерес поддержка семантических связей между авторами и публикациями, характеризующих вклад каждого автора в создание коллективных публикаций. Для этого таксономия связей должна включать классы связей с такой семантикой. Группой редакторов медицинских журналов совместно с Гарвардским университетом и медицинским благотворительным фондом Wellcome Trust из Лондона предложен *проект CRediT* [18, 19] открытого стандарта классификатора существенных для коллективно выполняемого исследования ролей его участников.

Нужно, наконец, упомянуть также модель научных данных *CERIF*, развиваемую европейской организацией euroCRIS (<http://www.eurocris.org/>). Один из

главных результатов развития CERIF – унифицированная концептуальная схема научно-исследовательской среды, называемая ее авторами *полной моделью данных (Full Data Model)* [20]. Эта модель рассматривается как единая основа разработок информационных систем (*Current Research Information Systems, CRIS*) для поддержки научно-организационной деятельности в разных странах и в различных научных организациях. Использование стандартизированной концептуальной схемы обеспечивает интероперабельность таких систем. В последнее время в развитии CERIF уделяется серьезное внимание проблеме спецификации семантики полной модели данных. В частности, предложена *онтология* [21], определяющая систему терминов для обозначения сущностей этой модели и отношения между ними.

В таксономии семантических связей, реализованной в системе Соционет, использованы указанные таксономии или их фрагменты. Включен также ряд дополнений, например, классы связей «пользователь–публикация», характеризующих рекомендации и информацию для автора публикации, или «пользователь–связь», характеризующих оценки семантики установленных связей пользователями системы. Подробнее эта таксономия обсуждается в следующем разделе.

2. ТАКСОНОМИЯ СЕМАНТИЧЕСКИХ СВЯЗЕЙ СИСТЕМЫ СОЦИОНЕТ

Работа с собственными публикациями и публикациями других исследователей является одним из важнейших элементов научной деятельности. В процессе этой работы возможны различные *реакции*, которые мотивируются, в частности, желанием исследователя определить и предоставить научному сообществу:

- собственные мнения о соотношении научных результатов, обсуждаемых в некоторых публикациях, с результатами его собственных работ или работ иных авторов;
- свои оценки содержания публикаций, доступных в среде системы;
- информацию о взаимосвязях между версиями своих публикаций и/или их компонентами;
- свои мнения об оценках других исследователей относительно представленных в системе публикаций;
- рекомендации авторам прочитанных публикаций, представленных в системе, направленные на развитие и/или улучшение этих публикаций;
- оценки своего вклада в создание коллективных публикаций;
- свои комментарии, аннотирующие фрагменты собственных публикаций или публикаций других авторов.

Оперативно и достаточно комфортно «материализовать» результаты таких реакций позволяет научная электронная библиотека, предоставляющая пользователям механизмы создания, публикации и использования семантических связей между информационными объектами ее контента. Каждая из указанных реакций *реализуется путем декларации семантической связи* с соответствующей семантикой. Результа-

ты такой деятельности становятся доступными научному сообществу пользователей библиотеки. Именно исходя из необходимости обеспечения средств для представления результатов перечисленных возможных реакций исследователя – пользователя системы, была определена описываемая таксономия семантических связей информационных объектов.

Таксономия семантических связей, реализованная в системе Соционет, имеет двухуровневую организацию и организована в виде набора контролируемых словарей. *Контролируемым словарем* называется некоторый упорядоченный набор допустимых значений, предоставляемый пользователю компьютерной системы для выбора конкретного значения, вводимого в окно экранной формы, с которым ассоциирован данный словарь. В нашем случае каждому классу верхнего уровня таксономии соответствует некоторый контролируемый словарь, а классам второго уровня – значения из этого словаря, представляющие имена классов, являющихся подклассами соответствующего класса верхнего уровня.

Для поддержки рассмотренных выше возможных реакций пользователей системы таксономия семантических связей системы Соционет включает следующий набор контролируемых словарей:

- Словарь видов развития и дополнений результатов исследований;
- Словарь видов профессиональных оценок публикаций;
- Словарь отношений между компонентами и версиями публикаций;
- Словарь видов мнений о существующих связях;
- Словарь видов близости публикаций;
- Словарь видов вкладов авторов в создание коллективных публикаций.

В системе используется также вспомогательный словарь связей без семантики, определенной в явном виде.

При аннотировании абстрактов публикаций, представленных в Соционет, пользователь ограничивается *только комментарием* к аннотируемому фрагменту абстракта. Поскольку аннотации представляются в системе также в виде семантических связей между профилем пользователя системы, осуществляющим аннотирование, и данной публикацией, в таком случае создается связь с неопределенным классом таксономии. Для создания таких связей используется специальный словарь таксономии с единственным классом связей «Семантика не определена».

Кроме указанных словарей, рассматриваемая таксономия включает также словари, необходимые для представления в системе административных связей между сотрудниками и организациями, а также между сотрудниками организации, связей авторства между сотрудниками/организациями и их публикациями, которые в данной статье не рассматриваются. Эти связи поддерживаются в системе для отображения научно-организационной структуры организации и генерации некоторых связанных с ней наукометрических показателей.

Важно отметить, что связи каждого класса таксономии можно декларировать только *на определенном множестве пар типов* исходного и целевого объекта связи. При создании конкретной связи, после того

как пользователь укажет исходный и целевой ее объекты, система использует *матрицу допустимости* классов связей, определяющую для каждой пары типов связываемых объектов возможные классы семантических связей между ними и предоставляет пользователю возможность выбора нужного класса связи из соответствующих контролируемых словарей.

В Соционет имеются средства создания, модификации и использования контролируемых словарей классов семантических связей для декларации семантических связей в системе, их публикации и изменения их свойств. Каждый словарь представлен в системе как *коллекция* информационных объектов типа *metrics*, каждый из которых представляет некоторый подкласс связей класса, которому соответствует данный словарь.

3. КОНТРОЛИРУЕМЫЕ СЛОВАРИ КЛАССОВ СЕМАНТИЧЕСКИХ СВЯЗЕЙ СОЦИОНЕТ

Контролируемые словари, представляющие таксономию связей в системе Соционет, предназначены для *классификаций* семантических связей (отношений), которые пользователь устанавливает между выбранными им информационными объектами контента системы. Каждый из словарей определяет набор классов, являющихся подклассами соответствующего этому словарю класса семантических связей таксономии. Классы связей представляются в словарях своими именами, отражающими семантику связей, относящихся к этому классу.

Для дальнейшего обсуждения необходимо уточнить некоторые используемые термины. Прежде всего, отметим, что в описании приведенных ниже контролируемых словарей термин *публикация* обозначает информационный объект системы, такой как книга, статья, научный отчет, тезисы доклада, автореферат или полный текст диссертации, препринт и т.п. Эти объекты могут быть следующих типов: *book* (книга), *article* (статья), *paper* (отчет, рабочая записка, тезисы доклада и др.), *news* (новость), *comment* (комментарий), *thesis* (диссертация), *artifact* (артефакт), *metrics* (класс связей).

Термин *персона* означает информационный объект, представляющий некоторое лицо (пользователя системы, администратора, автора публикации) и относящийся к типу *person*. Наконец, термин *организация* означает информационный объект, такой как вуз, научно-исследовательский институт или другое учреждение. Объекты, представляющие организации в Соционет, имеют тип *institution*.

В соответствии с технологией открытых архивов публикации, персоны, организации и другие информационные объекты представляются в системе Соционет их *описаниями*, содержащими для публикаций и других «материализованных» объектов ссылки на них в среде Веба, где они хранятся. Описания организаций и персон называются их *профилями*. Семантические связи между информационными объектами представляются в системе связями между их описаниями.

Приведенное ниже описание каждого контролируемого словаря, представляющего используемую в

системе Соционет таксономию семантических связей, включает его название (имя соответствующего класса верхнего уровня таксономии), общее описание семантики принадлежащих ему классов связей второго уровня и набор их имен, область его возможного применения.

3.1. Словарь видов развития и дополнений результатов исследований

Семантика: этот словарь включает классы связей, описывающих варианты использования результатов *целевых публикаций* в исходных, в частности, для научного вывода новых результатов.

Применение: классы связей из этого словаря могут использоваться только для связей вида «публикация–публикация».

Классы связей:

- Детализирует идею/метод/модель из
- Обобщает идею/метод/модель из
- Развивает результаты из
- Уточняет результаты из
- Исправляет ошибки в
- Опровергает результаты из
- Анализирует результаты из
- Интерпретирует результаты из
- Иллюстрирует результаты из
- Реализует идею/метод/модель из
- Использует метод из
- Использует модель из
- Использует данные из
- Использует статистику из
- Использует программное обеспечение из.

3.2. Словарь видов профессиональных оценок публикаций

Семантика: словарь включает классы связей, описывающих варианты оценки *целевых публикаций*, которые отражают мнение о них пользователей системы, профили которых служат исходными объектами таких связей.

Применение: классы связей из этого словаря могут использоваться только для связей вида «персона–публикация».

Классы связей:

- Рассматривает как наилучшую, наиболее релевантную работу по теме
- Подозревает плагиат
- Негативно оценивает целевую
- Считает, что использован ненаучный подход к получению результата
- Рассматривает как новаторский результат
- Рассматривает как очень интересный результат
- Считает поворотным пунктом для развития науки
- Позитивно оценивает целевую
- Рассматривает как результат с возможным опасным влиянием
- Полагает, что результат основан на заблуждении.

3.3. Словарь отношений между компонентами и версиями публикаций

Семантика: словарь включает классы связей, характеризующих отношения публикаций – исходных участников связей (абстрактов, оглавлений, предисловий и других компонентов публикаций, их версий, презентаций, родственных публикаций и др.) с целевыми публикациями.

Применение: классы связей из этого словаря могут использоваться только для связей вида «публикация–публикация».

Классы связей:

- Авторская версия (рукопись) для
- Версия с открытым доступом для
- Версия с небольшими изменениями для
- Версия с несущественными изменениями для
- Существенно переработанная версия для
- Идентичная копия для
- Презентация к
- Раздел, часть, глава из
- Абстракт (аннотация) для
- Новая редакция для
- Оглавление из
- Предисловие (введение) из
- Список литературы из.

3.4. Словарь видов мнений о существующих связях

Семантика: этот словарь включает классы связей, характеризующих мнение персон о семантике оценочных и других связей, выступающих в качестве целевых в данных связях.

Применение: классы связей из этого словаря могут использоваться только для связей вида «персона–связь».

Классы связей:

- Согласен с семантикой связи
- Не согласен с семантикой связи
- Считает неверно понятой оцениваемую/используемую работу
- Считает ошибочной оценку, представленную связью
- Считает завышенной оценку, представленную связью
- Считает заниженной оценку, представленную связью.

3.5. Словарь видов близости публикаций

Семантика: этот словарь включает классы связей публикаций создателя связи с рассматриваемой публикацией. Семантика создаваемых связей предоставляет авторам целевых публикаций полезные для их улучшения сведения, основанные на содержании выбранной создателем связи собственной публикации (исходной публикации связи). Связь этого класса позволяет независимо от связей цитирования целевой публикации с определенными для них мотивами цитирования сообщить ее автору сведения о развитии представленных в ней результатов или о результатах, близких к ним, обсуждаемых в исходной публикации связи.

Применение: классы связей из этого словаря могут использоваться только для связей вида «публикация–публикация».

Классы связей:

- У меня обсуждается близкая проблема
- У меня получены те же результаты
- У меня анализируются ваши результаты
- У меня иллюстрируются ваши данные/метод/модель/результаты
- У меня интерпретируются ваши результаты
- У меня детализируются ваши идея/метод/модель/результаты
- У меня обобщаются ваши идея/метод/модель/результаты
- У меня реализуются ваши идея/метод/модель
- У меня лучше данные/метод/модель
- В вашем случае применимы использованные мною программные средства
- У меня указаны и исправлены ваши ошибки
- У меня опровергаются ваши результаты.

3.6. Словарь видов вкладов авторов в создание коллективных публикаций

Семантика: словарь включает классы связей, характеризующих вклад персон-соавторов в подготовку коллективных публикаций – целевых объектов связи. При этом персона-соавтор выступает в качестве исходного участника связи.

Применение: классы связей из этого словаря могут использоваться только для связей вида «персона–публикация».

Классы связей:

- Администрирование проекта
- Научное руководство
- Компьютерные работы
- Проведение экспериментов
- Проведение исследования
- Концептуализация проблемы
- Курирование данных
- Разработка методологии
- Обеспечение финансирования
- Визуализация результатов. Подготовка начального варианта текста
- Критический анализ/доработка текста
- Обеспечение ресурсами
- Формальный анализ данных.

3.7. Словарь связей, семантика которых не представлена в явном виде

Семантика: это – особый словарь, содержащий классы связей между информационными объектами системы, семантика которых явным образом не представлена, а определяется по умолчанию и «встроена» в системные механизмы. В настоящее время этот словарь включает единственный класс. Описания связей, относящихся к этому классу, могут в дальнейшем обогащаться путем отнесения их к одному из классов связей описанных выше словарей и тем самым явного определения их семантики.

Применение: в системе Соционет класс из этого словаря используется для связей вида «персона–

публикация» со встроенной семантикой авторства, «персона–организация» с семантикой, характеризующей персон как сотрудников данных организаций, связей вида «публикация–публикация» для представления «немых» связей цитирования [2] и для некоторых других связей.

Классы связей:

- Семантика явным образом не определена.

3.8. Пример семантической связи

Предлагаемый ниже пример семантической связи между вымышленными публикациями использует класс семантической связи из контролируемого словаря разд. 3.1:

- *Исходная публикация связи:* Петров П.П. Оценка уровня инфляции в российской экономике. – М.: Экономика, 2015. – 200 с.

- *Класс связи:* Анализирует результаты из

- *Целевая публикация:* Васильев В.В. Промышленность и сельское хозяйство России в 2010–2014 гг. // Экономическая теория. – 2014. – № 4. – С. 12–25.

Приведенный выше набор контролируемых словарей семантических связей, поддерживаемых в системе Соционет, может пополняться новыми словарями, словари могут дополняться новыми классами, и имена классов в словарях могут изменяться.

4. СОЗДАНИЕ СЕМАНТИЧЕСКИХ СВЯЗЕЙ

Возможность создания семантических связей в системе Соционет доступна только для ее *авторизованных пользователей*. Это связано прежде всего с тем, что в описании создаваемой связи должен содержаться уникальный системный идентификатор пользователя – ее создателя. Поэтому при создании связи некоторым пользователем система должна его идентифицировать. Это возможно только при условии, если пользователь был зарегистрирован в системе и авторизовался в начале своего сеанса работы. Кроме того, идентификация пользователя необходима системе и в некоторых других случаях при создании связей. Например, для определения функций, доступных пользователю, в некоторых случаях система должна установить, является ли он автором рассматриваемой публикации (см. ниже). Но это невозможно, если пользователь не авторизовался в начале своего сеанса работы.

Идентификация создателя семантической связи необходима и потому, что факт существования связи и ее семантика становятся в системе *публично доступными*, и отсутствие анонимности автора повышает его ответственность перед научным сообществом за публикуемые мнение или оценку, представляемые создаваемой им связью.

В системе Соционет в настоящее время реализовано несколько способов создания семантических связей. Один из них основан на использовании специального «универсального» сервиса создания связей. Второй способ предназначен для использования пользователями-авторами представленных в системе публикаций. Третий способ служит для создания семантических связей путем семантического «обогащения» связей цитирования публикаций, представ-

ленных в списках использованных источников. Он также может использоваться только авторами цитирующих публикаций.

Описание каждой семантической связи, созданной пользователем системы, запоминается в указанной им его коллекции информационных объектов в среде хранения данных системы. Коллекции информационных объектов в системе представляют собой совокупность описаний однотипных информационных объектов.

Рассмотрим кратко указанные способы создания семантических связей.

Использование «универсального» сервиса создания связей. С помощью этого сервиса авторизованный пользователь системы может в онлайн-режиме создать семантическую связь любого допустимого класса между представленными в системе или «внешними» (не содержащимися в ее контенте, но доступными в Вебе) публикациями, за исключением семантических связей цитирования. Создание связей цитирования возможно только для автора цитирующей публикации. Для создания связи пользователь выбирает публикации, которые станут ее участниками. В зависимости от типов представляющих их информационных объектов ему доступны для выбора некоторые контролируемые словари классов связей. Пользователь выбирает подходящий для его цели словарь и один из входящих в этот словарь класс связей, тем самым определяя семантику создаваемой связи. Кроме того, пользователь – создатель связи может задать некоторый комментарий, входящий в описание связи. Наконец, пользователь задает также локальный *идентификатор создаваемой связи* – идентификатор, уникальный в рамках той коллекции информационных объектов, в которую он намерен включить эту связь.

В результате для создаваемой связи создается ее *описание* [1, 6], включающее: глобальный (в рамках всей системы) уникальный идентификатор, генерируемый системой на основе заданного пользователем локального идентификатора; глобальные системные идентификаторы публикаций – участников связи; идентификатор словаря (класса) таксономии, к которому относится создаваемая связь; если необходимо, комментарий создателя связи. В описании указывается также идентификатор автора связи. В случае использования «внешней» публикации в качестве участника связи, она должна быть доступна в Вебе, и в качестве ее глобального идентификатора используется URI этой публикации. Комментарий в описании связи может использоваться для неформального описания некоторых ее свойств или причин, побудивших автора связи ее создать.

Использование специального интерфейса на странице описания публикации. Этот способ более прост в использовании для пользователя системы по сравнению с использованием «универсального» сервиса. Для того чтобы им воспользоваться, нужно вызвать на экран описание публикации, которая станет *целевым участником* создаваемой связи. Доступ к рассматриваемому интерфейсу осуществляется кликами на активных областях («виртуальных кнопках») этого экрана и возможен как для автора данной публикации, так и для ее читателя.

При использовании рассматриваемого интерфейса поведение системы и доступные пользователю возможности зависят от того, какую роль по отношению к данной публикации играет пользователь. Когда авторизованный пользователь просматривает описание некоторой публикации, система опознает его как ее автора или как читателя. Пользователь квалифицируется как ее *автор* в случае, если идентификатор этого пользователя указан как значение соответствующего атрибута в описании этой публикации. В остальных случаях она рассматривает его как читателя.

Если пользователь выступает в роли автора, исходным объектом создаваемой связи может являться его персональный профиль, а целевым объектом – его публикация (ее описание). При этом автор может указать свой *вклад в подготовку этой публикации* (используя словарь из разд. 3.6), если она – коллективная. Он может также *аннотировать* абстракт своей публикации (используя словарь из разд. 3.7) с целью актуализации или уточнения его содержания.

Кроме этого, пользователь-автор может установить семантическую связь между парой *своих* публикаций. В этом случае он имеет возможность использовать словарь *видов развития и дополнений* (разд. 3.1) для того, чтобы указать подходящее отношение между изложенными в них результатами исследований, либо указать характер отношений *между различными версиями* или компонентами одной и той же своей публикации (разд. 3.3).

Пользователь, выступающий в роли *читателя*, может устанавливать связи от представляющего его персонального профиля (исходный объект) к публикациям других авторов (целевые объекты). Таким способом, просматривая описание выбранной публикации, он может *аннотировать* фрагменты ее абстракта, при желании дать *профессиональную оценку* такой публикации (разд. 3.2), которая станет целевой в создаваемой семантической связи.

Пользователь-читатель может также создать связь от выбранной своей публикации (исходный объект) к чужой публикации (целевой объект). Это могут быть, например, связи *информационного характера* (разд. 3.5), с помощью которых автор исходной публикации сообщает автору целевой публикации *о близости их содержания*. Могут быть также установлены связи классов из словаря *видов развития и дополнений* (разд. 3.1) от выбранной публикации читателя к публикации, описание которой просматривается на экране.

Семантическое обогащение связей цитирования. Третий способ создания семантических связей в системе Соционет относится к связям между данной публикацией как исходной и публикациями, указанными в пристатейном списке литературы. По существу, речь идет о семантическом обогащении «*немых*» связей цитирования, уже включенных ранее в описание публикации. «*Немыми*» мы называем традиционные связи цитирования, поскольку они не несут какой-либо информации о мотивах, побудивших автора цитировать ту или иную публикацию. Для таких связей в результате определяется явно указанная семантика на основе таксономии связей, поддерживаемой в системе. Такие связи может создавать толь-

ко пользователь – автор цитирующей публикации, после того как он авторизовался при входе в систему.

Для создания таких связей пользователь должен вызвать на экран описание цитирующей публикации. Если этот пользователь был объявлен как ее автор, система предоставит ему возможность определения семантики связей цитирования. У каждой из публикаций списка использованных источников цитирующей публикации на экране появляется виртуальная кнопка, кликнув на которую пользователь получает доступ к словарю *видов развития и дополнений* (разд. 3.1). Выбрав подходящий класс для создаваемой связи, пользователь тем самым определяет характер использования цитируемого источника в этой своей работе и порождает соответствующую связь, которая запоминается в указанной им коллекции информационных объектов системы.

5. ИСПОЛЬЗОВАНИЕ СЕМАНТИЧЕСКИХ СВЯЗЕЙ В СОЦИОНЕТ

Поддерживаемые в Соционет семантические связи между ее информационными объектами обеспечивают ряд новых возможностей по сравнению с обеспечиваемыми в традиционных электронных библиотеках. Ограничимся здесь кратким обсуждением наиболее важных из этих возможностей.

Семантическое структурирование корпуса научных знаний в библиотеке. Прежде всего, явная декларация в системе семантических связей между публикациями позволяет семантически структурировать корпус научных знаний, содержащихся в системе [2, 3]. Формирование семантических связей осуществляется децентрализованно авторизованными пользователями в онлайн-режиме, и в этой деятельности система поддерживает взаимодействие пользователей, обеспечивая им возможности для реагирования на создание связей другими пользователями. Стимулирует ответные реакции пользователей на создание семантических связей, участниками которых являются их публикации, предусмотренный в системе сервис оповещения о таких событиях.

Визуализация связей и семантическая навигация. При обращении пользователя к какой-либо из имеющихся в системе публикаций, он имеет возможность визуально анализировать ближайшую окрестность узла семантической сети, соответствующего данной публикации, т.е. данный узел и все непосредственно связанные с ним узлы с указанием классов связей. При просмотре описания публикации при желании можно видеть все входящие и исходящие связи данного узла. Это позволяет получать информацию о месте публикации в корпусе научных знаний, о соотношении этой публикации с другими, представленными в системе, о ее оценках пользователями системы. Вместе с тем, визуализация связей каждого узла семантической сети дает возможность пользователю осуществлять *семантическую навигацию* по сети, пошагово перемещаясь от одной публикации к другой по связям интересующих его классов.

Пользуясь визуализированным представлением окрестности каждого неизолированного узла семантической сети, можно легко найти иные версии публикации, соответствующей этому узлу, ее разделы

или презентации, если они также представлены в системе. Вместе с тем можно отфильтровать слой сети, соответствующий интересующему пользователя классу связей, например, связей, определяющих характер использования результатов, представленных в одних публикациях, – в других. Далее, стартуя от текущего узла сети, можно навигацией по полученной подсети проследить эволюцию научных результатов в некоторой области науки, выявить относящиеся к ней основополагающие работы. Конечно же, все это полезно при условии достаточно представительного контента системы как относительно состава представленных в нем публикаций, так и относительно семантических связей между ними.

Используя подсеть оценочных связей, можно найти работы либо мнения конкретных пользователей системы, в которых данная публикация подвергается критике или, наоборот, получила положительные оценки.

Наукометрические исследования. Другое важное направление использования семантических связей между информационными объектами системы – это использование семантической сети, построенной на ее контенте, как источника данных для наукометрических исследований, более информативного по сравнению с традиционно используемыми на практике [9]. Традиционные наукометрические исследования обычно основаны на использовании связей цитирования в пристатейных списках публикаций из периодики. Однако эти связи не несут никакой информации о мотивах цитирования в отличие от семантически обогащенных связей цитирования в системе Соционет. Кроме того, в наукометрии могут быть использованы и поддерживаемые в системе связи, отличные от связей цитирования, имеющие разную семантику.

Наукометрические сервисы системы Соционет генерируют статистические данные дифференцированно по классам связей, в том числе и для связей цитирования. Такие данные позволяют более глубоко анализировать и оценивать научное качество представленных в системе публикаций, степень их влияния на развитие научных исследований, изучать содержательные наукометрические «портреты» как отдельных ученых, так и исследовательских организаций в целом, получать более осмысленные оценки их научной продуктивности. Использование в системе семантических связей, отражающих вклад автора в создание коллективной публикации, позволяет также проводить наукометрические измерения, связанные с персональными оценками этого аспекта научной деятельности для отдельных исследователей.

Научные коммуникации. Нужно, наконец, упомянуть еще одно достаточно существенное направление использования семантических связей в системе [4]. Создавая семантические связи, авторизованные пользователи могут с использованием структурированных (имен классов связей) и/или неструктурированных данных (комментариев в описаниях создаваемых связей) публиковать в системе свои профессиональные оценки публикаций других авторов, предлагать им рекомендации по развитию конкретных их работ, комментировать абстракты своих

и чужих публикаций. Наряду с этим имеется возможность публиковать мнения и оценки относительно оценок, высказанных другими пользователями. Для этого нужно создавать связи собственного профиля со связями, выражающими оцениваемые оценки. Таким образом, на полигоне системы могут поддерживаться альтернативные формы рецензирования научных публикаций, научные дискуссии и обмен научной информацией, ассоциированной с представленными в системе публикациями. Авторы публикаций и других научных информационных объектов, которые становятся участниками создаваемых связей, информируются об этом по электронной почте сервисом уведомления, имеющимся в системе Соционет. Получение оповещения стимулирует получателя реагировать на него, создавая соответствующую новую связь, и тем самым поддерживать такую дискуссию. Все созданные в описанных процессах семантические связи включаются в пространство наукометрических измерений в системе с учетом их классов в таксономии.

ЗАКЛЮЧЕНИЕ

Описанная в статье таксономия семантических связей реализована в системе Соционет. Созданы механизмы формирования, поддержки и использования контролируемых словарей, а также декларируемых с их помощью семантических связей. Эти механизмы позволяют при необходимости расширять и корректировать поддерживаемую в системе таксономию семантических связей, не затрагивая программного обеспечения системы. Разработаны средства контекстной визуализации семантических связей для узлов семантической сети, а также ряд других пользовательских сервисов системы. Все эти средства находятся в стадии опытной эксплуатации.

СПИСОК ЛИТЕРАТУРЫ

1. Когаловский М.Р., Паринов С.И. Технологии социальной сети для создания семантических связей информационных объектов в научной электронной библиотеке // Программирование. МАИК/Наука «Интерпериодика». – 2014. – Т. 40. – № 6.
2. Когаловский М.Р., Паринов С.И. Семантическое структурирование контента научных электронных библиотек на основе онтологий // Современные технологии интеграции информационных ресурсов: сборник научных трудов. – СПб.: Президентская библиотека им. Б.Н. Ельцина, 2011. – С. 26-45.
3. Паринов С.И., Когаловский М.Р. Технологии семантического структурирования контента научных электронных библиотек // Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2011. Воронеж, 19–22 октября 2011 г.». – Воронеж: Воронежский государственный университет, 2011. – С. 94–103.

4. Когаловский М.Р., Паринов С.И. Научные коммуникации в среде семантически обогащаемых электронных библиотек // Программная инженерия. – 2015. – № 4. – С. 31–38.
5. Kogalovsky M.R., Parinov S.I. Social Network Technologies for Semantic Linking of Information Objects in Scientific Digital Library // Programming and Computer Software. – 2014. – Vol. 40, № 6. – P. 313–320.
6. Когаловский М.Р., Паринов С.И. Новый источник данных для наукометрических исследований // Труды XV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2013» (Ярославль, 14–17 октября 2013 г.). – Ярославль: Ярославский университет, 2013. – С.107–117.
7. Parinov S. Open Repository of Semantic Linkages // Proceedings of 11th International Conference on Current Research Information Systems e-Infrastructure for Research and Innovations (CRIS 2012), Prague. – 2012. – URL: <http://socionet.ru/publication.xml?h=repec:rus:mqijxk:29>.
8. Parinov S. Towards a Semantic Segment of a Research e-Infrastructure: necessary information objects, tools and services // Metadata and Semantics Research, Communications in Computer and Information Science / eds. J.M. Doderо, M. Palomo-Duarte, P. Karampiperis. Springer, 2012. – Vol. 343. – P. 133–145. – URL: <http://socionet.ru/pub.xml?h=RePEc:rus:mqijxk:30>.
9. Parinov S., Kogalovsky M. Semantic Linkages in Research Information Systems as a New Data Source for Scientometric Studies // Scientometrics. – 2014. – Vol. 98, Issue 2. – P. 927–943.
10. Parinov S. Semantic enrichment of research outputs metadata: new CRIS facilities for authors // MTSR 2014, 8th Metadata and Semantics Research Conference, 27–29 November 2014, Karlsruhe, Germany.
11. Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайн-овых сервисов // Российский научный электронный журнал «Электронные библиотеки». – 2003. – Том 6. – Вып. 1. – URL: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part1/PLP>.
12. Open Archives Initiative. – URL: <http://www.openarchives.org/>.
13. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2015-01-08. – 2015. – URL: <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>.
14. Shotton D. Open Citations and Related Work. Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. October 14, 2010. – URL: <http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/>.
15. Shotton D., Peroni S. Semantic annotation of publication entities using the SPAR (Semantic Publishing and Referencing) Ontologies / Beyond the PDF Workshop, La Jolla, 19 January 2011. – URL: <http://speroni.web.cs.unibo.it/publications/shotton-2010-semantic-annotation-publication.pdf>.
16. Semantic Web Applications in Neuromedicine (SWAN) Ontology. W3C Interest Group Note, 20 October 2009. – URL: <http://www.w3.org/TR/hcls-swan/>.
17. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. – URL: <http://www.w3.org/TR/skos-reference/>.
18. Liz Allen, Amy Brand, Jo Scott, Micah Altman, Marjorie Hlava. Credit where credit is due // Nature / International weekly journal of science. – 2014. – Vol. 508, Issue 7496, April. – URL: http://www.nature.com/polopoly_fs/1.15033!/menu/main/topColumns/topLeftColumn/pdf/508312a.pdf.
19. CRediT. An open standard for expressing roles intrinsic to research. Taxonomy. – 2014. – URL: <http://credit.casrai.org/proposed-taxonomy/>.
20. CERIF 1.3 Full Data Model (FDM): Introduction and Specification. euroCRIS, 2012. – URL: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf.
21. CERIF 1.3 Semantics: Research Vocabulary. CERIF Task Group, euroCRIS, 2012. –URL: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_Semantics.pdf.

Материал поступил в редакцию 10.03.15.

Сведения об авторах

КОГАЛОВСКИЙ Михаил Рувимович – кандидат технических наук, доцент, зав. лабораторией Института проблем рынка РАН, Москва.
e-mail: kogalov@gmail.com

ПАРИНОВ Сергей Иванович – доктор технических наук, зам. директора Центрального экономико-математического института РАН, Москва.
e-mail: sparinov@gmail.com

Лингвистические основания путеводителя по вопросам сохранности документов

Рассматривается путеводитель по электронным ресурсам в тематической области сохранения документов. Показывается, как на разных этапах создания сетевого ресурса производится декомпозиция его лингвистического материала, состоящая в разделении общих понятий на более узкие, и путеводитель «выстраивается» из некоторых общих понятий. Описываются преимущества вышесказанной логики построения путеводителя.

Ключевые слова: путеводитель, веббиблиография, сохранность документов, консервация документов, электронные библиографии, БАН, фасетный анализ

Предметные путеводители, или коллекции ссылок по каким-либо темам (они называются также *webliographies* и *web guides*) становятся все более распространенным и привычным для пользователей инструментом информационного поиска. В тематической области «консервация документов» путеводители по сетевой информации имеются в большом количестве (более 60), и в том числе – информационные ресурсы [1–6]. В Библиотеке Академии наук (БАН) разработан первый российский путеводитель по этой тематике, доступный по электронному адресу <http://91.151.182.200:8083> и описываемый далее в связи с декомпозицией содержания тематических рубрик и запросов, используемых для комплектования путеводителя. Декомпозиция в этой работе описывается как основной лингвистический инструмент построения обсуждаемого сетевого ресурса.

1. ДЕКОМПОЗИЦИЯ ЗАПРОСОВ В ДОКУМЕНТАЛЬНЫХ СИСТЕМАХ

Первое упоминание о декомпозиции как об особом приеме обработки информационных запросов имеется у Р.П. Вчерашнего [7–8]. Декомпозиция в его освещении служит для повышения полноты результатов информационного поиска по индексам Международной классификации изобретений, при работе с которой строятся сложные поисковые выражения. Так как при поисках по МКИ и при проведении патентных исследований особенно нежелательны информационные потери, в поисковые выражения запросов включаются не только рубрики МКИ, прямо обозначающие тему запроса (например, *печи туннельные*), но также рубрики нижестоящие или ассоциативные по отношению к ним. Данные рубрики могут называть, например, конструктивные части

устройств – *системы управления, устройства для садки и съема изделий, рекуператоры*, т.е. устройства, которые могут быть применимы в туннельных печах. Необходимость в использовании декомпозиции запросов определяется также особенностями самих изобретений как объектов патентования. «Объект техники, как правило, содержит совокупность частных технических решений и поэтому может быть заиндексирован только совокупностью индексов МКИ. Для того чтобы заиндексировать разработку по МКИ, ее необходимо «раздетализировать» до отдельных частей технических решений, т.е. произвести декомпозицию объекта» [9, с. 184–185]. Такая декомпозиция чаще всего производится по трем основаниям или фасетам: конструктивному, выполняемой функции и составу материалов.

Под декомпозицией мы имеем в виду операцию разделения общих запросов на несколько связанных между собой более частных запросов. Так, занимаясь исследованием информации и написанием текста по теме «Разработка тезауруса по сохранности документов», автор данной статьи расчленил эту общую тему на несколько частных подтем, таких как *принципы составления словника, нормализация лексики, обработка синонимов, обработка родовидовых отношений* и др., и собирал информацию по этим проблемам. Декомпозиция в данном случае облегчала работу с литературой, в большей части сводившейся к компиляции собранных литературных источников.

Употребление декомпозиции запросов особенно эффективно тогда, когда на запрос выдается очень большое количество информации. Так, в ответ на запрос «Огнеупорные волокна» (цель поиска – написание обзора литературы для докторской диссертации) в системе по огнеупорам было выдано около 3 тыс. документов. Поскольку предоставление пользовате-

лю всей такой информации без ее дополнительной обработки представлялось неправильным из-за ее большого объема, была сделана декомпозиция запроса, соответствующая содержанию глав диссертации и построенная вместе с пользователем таким образом: *карбидкремниевые волокна, нитридкремниевые волокна, окисные волокна, способы производства волокон, методы изучения волокон* и т.п. При проведении декомпозиций запросов иногда применялась операция логического отрицания, устранявшая повторную выдачу документов по подзапросам. Целесообразность использования операции «без» в таких случаях обязательно согласовывалась с пользователем, так как для некоторых пользователей неоднократная, но обеспечивающая многоаспектное освещение темы запроса выдача документов является более предпочтительной.

В других ситуациях декомпозиции запросов могут быть неоправданными и могут приводить к нежелательным результатам. В начальный период работы системы по огнеупорам для системы ИРИ (избирательное распределение информации) весьма характерными являлись следующие серии запросов, хорошо раскрывающие индивидуальные информационные потребности каждого конкретного пользователя и полученные с помощью декомпозиции его более общей информационной потребности:

Абонент А

Магнезиальные огнеупоры для конвертеров
Магнезиально-углеродистые огнеупоры для конвертеров
Огнеупоры для конвертеров с донной продувкой
Факельное торкретирование конвертеров
Устройства для торкретирования конвертеров
Конструкция футеровок конвертеров

Абонент Б

Конструкция скользящих затворов (шиберных и поворотных)
Плиты скользящих затворов
Шлифование плит скользящих затворов
Устройства для прижима плит скользящих затворов
Заращение скользящих затворов

Абонент В

Обжиг огнеупорного сырья
Обжиг огнеупорных изделий
Огнеупоры для туннельных печей
Автоматизация туннельных печей
Способы садки в туннельные печи
Конструкция садки
Методы расчета туннельных печей
Вспомогательное оборудование для туннельных печей
Кинетика спекания огнеупорных окислов и т.п.

Поскольку в БД обрабатывалась реферативная и, следовательно, многоаспектная информация, при большой глубине индексирования документов ключевыми словами такая декомпозиция запросов приводила к тому, что документы могли соответствовать одновременно двум или более информационным запросам, и поэтому регулярно имелась неоднократная выдача информации по разным запросам из серий. Данная выдача воспринималась как шумовая, и приведенные серии запросов с течением времени обобщались и сокращались до одного или двух подзапросов (конвертеры; скользящие затворы; обжиг, туннельные печи).

Декомпозиция запросов при поиске в Интернете в существующих рекомендациях никак не рассматривается. Но возможность такой обработки запросов вытекает из правила, по которому наиболее эффективным при проведении поиска являются узкие по содержанию запросы. Это правило может иметь различные формулировки и, в частности, следующие: «Правильный запрос состоит из нескольких слов, потому что по одному слову обычно трудно понять, что вы хотите найти» [10]; «Для достижения необходимого результата уточняйте (конкретизируйте) запрос, используя ключевые слова. Если вы ищете сведения по автомобилю Toyota, например, то ваш запрос должен быть: „автомобиль Toyota“, а не „автомобиль“. Если вас интересует творчество Льва Толстого, то введите в строку поиска, кроме имени писателя, еще и слово „творчество“ через пробел» [11]; «Формировать запрос надо максимально точно, используя все возможности механизма составления запроса. При точном формировании запроса процент информационного шума будет намного ниже» [12].

Из приведенного положения следует, что возможным путем обработки широких запросов является их представление в виде серии узких запросов, совокупное содержание которых более или менее равноценно содержанию подлежащего декомпозиции запроса. Ниже описаны стадии построения путеводителя, на которых запросы и рубрики представляются в виде нескольких подзапросов (подрубрик).

2. ФОРМИРОВАНИЕ КЛАССИФИКАЦИОННОЙ СХЕМЫ

В классификационной схеме путеводителя используются предметные рубрики, и в основании данной схемы лежит предметизационная модель информации, ранее широко применявшаяся в имеющейся в БАН библиографической базе данных по консервации документов [13–14]. В соответствии с этой моделью в содержании информации различают ее предмет и аспект. В роли предмета выступают слова со значением ‘документы’ и ‘материалы’, как *paper, cardboard, leather, parchment; books, bookbinding, photographs, manuscripts*. Аспекты ассоциируются с содержанием ‘консервация’ – *conservation, preservation, storage* – и нижестоящими (видовыми) по отношению к ним названиями. Существенным дополнением предметизационной модели в ее семантической интерпретации в БАН является введение в нее представления о «нулевом» выражении позиции предмета или аспекта, обозначаемого далее символом Ø [15]. Содержание этой позиции может быть различным в зависимости от условий, в которых она проявляется.

Рассматриваемые вне контекстного окружения и без всякой привязки к какой-либо области знания, нулевые позиции имеют значение ‘всякий’, ‘любой’ предмет или аспект информации. Таково, например, содержание *нулевого предмета* в тематической рубрике **Environment Control**, интерпретируемой как $S[Ø] - P[Environment Control]$. Здесь нулевой предмет информации означает ‘все об’ *Environment Control*, и ему соответствуют результаты информационного поиска в Интернете по рассматриваемой тематической рубрике. Поиск по рубрике *Environment*

Control дает большое количество шумовых веб-страниц о контроле среды, но применительно не к консервации документов, а к самой широкой экологической, сельскохозяйственной и медико-биологической проблематике. Нулевой аспект информации имеется в рубрике **Paper and Cardboard**. Он хорошо выявляется при сравнении этой рубрики с нижестоящими по отношению к ней (видовыми) подразrubиками, каковыми являются: Types of paper and paperboard (S [Paper and Cardboard] – P [Types]), Paper & Cardboard Manufacture (... – P [Production]), Paper Deacidification (... – P [Deacidification]), Paper aging (... – P [Aging]). Приведенные примеры показывают, что содержание рубрики **Paper and Cardboard** следует интерпретировать, как S [Paper and Cardboard] – P [Ø], где нулевой аспект информации имеет значение 'любая возможная сторона рассмотрения' предмета.

В контексте предметной области «консервация документов» нулевые (или, иначе сказать, актуальные и в явном виде не выраженные) предметы и аспекты информации имеют другое значение. Нулевой предмет информации имеет значение 'документы' или 'хранилища', которое раскрывают контексты, как S [Книги, Коллекции, Фонды] – P [Контроль окружающей среды], S [Библиотеки, Архивы, Музеи] – P [Контроль окружающей среды]. Нулевой аспект информации соответствует содержанию 'консервация, обеспечение сохранности', как в тематических рубриках *Рукописи* (... – P [Ø]), *Фотографии* (... – P [Ø]), *Кожа* (... – P [Ø]). В тематических рубриках путеводителя нулевые позиции вполне допустимы, так как их содержание уточняется тем, что данные рубрики действуют в рамках путеводителя по сохранности документов. Иная ситуация с запросами, на основе которых проводится комплектование путеводителя. При обработке запросов нулевые позиции должны эксплицироваться и в явном виде описываться ключевыми словами в поисковых выражениях запросов. Техника подготовки запросов и, в частности, экспликации «информационных нулей» описывается в разделе 4 настоящей статьи.

Из рассмотренных выше понятий (фасетов) формируется обобщенная схема предметной области, охватывающая ее основное предметно-аспектное содержание и имеющая следующий вид:

Кто есть кто
Бумага, картон, книги
Рукописи
Фотографии
Консервация
Климатические условия
Биологический фактор
Стихийные бедствия, аварии

За исключением рубрики *Кто есть кто* [в области обеспечения сохранности документов], классы в рассматриваемой схеме являются ближайшими видовыми названиями по отношению к терминам «документы» и «консервация». Содержание рубрик включает в себя также вышеописанные нулевые позиции – нулевой аспект информации (Бумага, картон, книги – P [Ø], Рукописи – P [Ø], Фотографии – P [Ø]) и нуле-

вой предмет сообщения или запроса (S [Ø] – Консервация, S [Ø] – Климатические условия, S [Ø] – Биологический фактор, S [Ø] – Стихийные бедствия, аварии). Получение из обобщенного представления области ее более детализованного описания ведется при помощи операции декомпозиции рубрик. Вышестоящие (родовые) названия рубрик разделяются на нижестоящие (видовые) подразrubики. Указанным образом получается описываемая ниже классификационная схема путеводителя по вопросам сохранности документов. Цифры в скобках обозначают количество сайтов, стоящих за рубриками:

People & Organizations (10)

Finding Conservation Organizations

Finding People in Conservation

Database of Conservators & Bookbinders in Practice

Paper and Cardboard (2160)

Types of paper and paperboard (1420)

Paper & Cardboard Manufacture (320)

Paper Acidity & Deacidification (170)

Paper Aging (250)

Paper & Book Conservation (2650)

Reference Information (230)

General & Overview Information (160)

Professional Standards (190)

Preventive Conservation (170)

Bookbinding (470)

Book Repair & Restoration (304)

Manuscripts & Rare Books Conservation (325)

Leather Conservation (480)

Photographs Conservation (307)

Environment Management (730)

Environment Control (250)

Indoor Air Quality, Dust (190)

Lighting, UV & IR Radiation (170)

Temperature & Humidity (120)

Pest Management (2820)

Pest Control (310)

Integrated Pest Management (110)

Mold & Mildew (880)

Mass Treatment (1110)

Biocides/Pesticides (410)

Disasters & Incidents (950)

Disaster Prevention & Management (230)

Disaster Mitigation & Preparedness (200)

Disaster Response & Recovery (100)

Fire Protection (80)

Floods (80)

Incidents, Theft, Vandalism (190)

Disaster resources (50)

Similar Web Resources (60)

Other Conservation Webguides in English

Многие классы имеют многоуровневую иерархию понятий. Класс «Бумага и картон», например, в развернутом виде включает в себя не 4, а 24 рубрики только первого уровня их подчиненности:

Paper and Cardboard (1760)

Types of paper and paperboard

Printing Paper

Archival (acid-free) Paper

Archival Cardboard & Archival boxes

Restoration Paper

Bookbinding Board

Other Papers

Paper & Cardboard Manufacture
[Printing Paper](#)
[Archival \(acid-free\) Paper](#)
[Archival Cardboard & Archival boxes](#)
[Restoration Paper](#)
[Bookbinding Board](#)
[Other Papers](#)

Paper Acidity & Deacidification
[Bibliographies](#)
[Dictionaries](#)
[Guidelines & Manuals](#)
[Paper Acidity](#)
[Paper Deacidification](#)
[Directories](#)

Paper Aging
[Bibliographies](#)
[Dictionaries](#)
[Natural Aging](#)
[Accelerated Aging](#)
[Imitation of Paper Aging](#)
[Directories](#)

Эти предметные рубрики, в свою очередь, разделяются на подрубрики, как *Printing Paper* => *Offset Paper*, *Newsprint*, *Paper for digital printing*, *Typographical paper (paper for high print)*, *Coated Paper*. Описанным образом и с применением нескольких уровней иерархии понятий формируется сложная и разветвленная понятийная схема путеводителя по сохранности документов. С учетом описанной иерархии понятий в классификационной схеме путеводителя представлены 234 рубрики, из них 73 – русские и 161 – английские. При каждой из этих рубрик имеется список ссылок на информацию по теме. Данные списки являются основным инструментом информационного поиска в обсуждаемом путеводителе.

3. ПОДГОТОВКА СПИСКОВ ТЕМАТИЧЕСКИХ ССЫЛОК

Предметизационная модель информации также использовалась для получения названных списков. В отличие от тематических рубрик, имеющих в классификационной схеме путеводителя (где предмет составляет понятие ‘документы’), в списках ссылок предметами служат слова, называющие организации или сайты, к которым относится ссылка. Аспектами рубрик выступают слова, называющие сообщаемое о предмете. Вышеописанное структурирование ссылок обеспечивает некоторые поисковые преимущества и может служить основанием для суждения о полезности ссылок.

Предметно-аспектное форматирование текстов ссылок может являться горизонтальным, как *Aarons Restoration – Mold Removal*, *Photo Gallery*, и вертикальным, когда к сайту относится несколько ссылок. В этом случае формируется папка, в которой наименование папки составляет предмет, а ее содержимое составляют аспекты:

Centers for Disease Control and Prevention
[Dampness and Mold in Buildings](#)
[Mold after a Disaster](#)
[Mold and Related Issues](#)
[Mold General Information, Basic Facts](#)
[Mold Links](#)

В ряде случаев информацию можно связать с двумя или более организациями, некоторые статьи, например, имеют несколько авторов, и данные авторы могут работать в разных местах. При подготовке ссылок на такие статьи для простоты описания информации предметом ссылки считается место работы только первого автора публикации. При выборе между названием сайта, на котором опубликована информация, и названием описываемой профильной организации предпочтение отдается последней. Таким образом была сформирована, например, папка ссылок:

Институт научной информации по общественным наукам (ИНИОН)

[Алтайские сыроделы предложили свою технологию для борьбы с плесенью книг](#)
[Книги библиотеки ИНИОН могут стать недоступными на десятилетия из-за грибка](#)
[Книги из сгоревшей библиотеки ИНИОН РАН появились на уличных развалах](#)
[Книжное пепелище: библиотеку ИНИОНа доедает грибок](#)
[Роман Бизюков о пожаре в библиотеке ИНИОН](#)
[Спасение библиотеки ИНИОН](#)
[Ученые нашли способ, как сберечь раритетные издания](#)
[Часть библиотеки ИНИОН заморозят](#)
[Что происходит с библиотекой ИНИОН](#)

Для специалистов по консервации документов не очень существенно место издания рассматриваемой информации, а именно сайты газеты «Аргументы и факты», [newsmsk.com](#), «Частный корреспондент», [vesti.ru](#), газеты «Метро» др. Более ценным обозначением предмета является наименование профильной организации (библиотеки) и информация о плесневых грибах в ИНИОН, которая в путеводителе агрегируется как аспект информации вышеуказанным образом.

Рубрики (декомпозиции рубрик) могут являться регулярными, применимыми ко всем тематическим классам, и иррегулярными, применимыми только к какому-то из них. Регулярные рубрики выделяются по признаку ‘жанрово-видовая характеристика информации’ и представлены в зарубежной части путеводителя шестью тематическими классами: библиографии, словари, базы данных, учебники и руководства, основные ссылки и директории. Иррегулярные декомпозиции рубрик не имеют какого-то одного семантического основания и могут быть самыми разными по содержанию. Так, «излишне широкая» по наполнению предметная рубрика *Mold* в англоязычной части путеводителя разделена на подрубрики [Mold on Paper, in Libraries](#), [Mold on Leather](#), [Mold in Homes: Mold Identification, Testing](#), [Mold in Homes: Mold Removal & Remediation](#), [Fungicides](#). В русской части путеводителя та же предметная рубрика (Плесень) декомпозирована несколько иначе, а именно [Плесень на бумаге и коже](#), [Контроль, идентификация плесени](#), [Борьба с плесенью](#) и [Фунгициды](#).

Ниже приводится декомпозиция списков ссылок по теме *Mold & Mildew*, где регулярные рубрики (папки ссылок) обозначены жирным шрифтом:

Mold & Mildew (880)

Bibliographies (20)

Dictionaries (25)

Databases (20)

Guidelines & Manuals (40)

Directories (20)

Mold on Paper, in Libraries (110)

Mold on Leather (80)

Mold in Homes: Mold Identification, Testing (140)

Mold in Homes: Mold Removal & Remediation (155)

Fungicides (80)

Abbey Newsletter

Abbey Publication – Mold as a threat to human health

ABC News – Is Hidden Mold at Home Making You Sick?

Acaem.com – Treatment of Mold Outbreaks

AECOM Canada – Mould Investigation and Assessment Report

Alameda County Public Health Department – Mold Resource List

Alaska State Museum – A Fungus Among Us: What's That White Stuff?

Alberta Health Services Mouldy Homes & Buildings

AllTech Environmental Services – Mold Resources

Angies List – Is a Mold problem in your Home making you Sick?

Amazon.com – Water in Buildings: An Architect's Guide to Moisture and Mold / by William B. Rose

American Institute for Conservation (AIC)

Angies List – Beware of Mold when buying a foreclosed home

Answers.com – Houses with Mold Do Not Sell Quick in the Real Estate Market

Asthma and Allergy Foundation of America – Mold Allergy

Asthma UK – Moulds & Fungi

Bartlett Heating & Cooling – Mold Resources

Berkeley Natural History Museums – Life History and Ecology of the Fungi

Black Mold Resources – Indoor Mold Resources

Black Country Living Museum – Mould & Pest

Management Day Course

Blue Page.org – Toxic Mold

British Listed Buildings – Listed Buildings in Mold,

Flintshire, Wales

В разделе **Bibliographies** содержатся ссылки на полнотекстовые библиографические указатели и запросы по теме предметной рубрики для проведения по ним поисков информации в базах данных. Информация в этом разделе представлена следующим образом:

Bibliographies (20)

Amazon.co.uk: Mushrooms & Fungi Books

Australian National Herbarium – Fungi: Reference Information

Cornell University – Mycology Resources: Collections

Forager Press – Books about Mold

Inspectapedia – Mold bibliography

Mold Remediation Guidelines – Notes and References [bibliography]

Mold Survivor – The largest library of mold-related scientific literature [Web Links]

MykoWeb – A Bibliography of Mushrooms & Other Fungi

National Archives and Records Administration (NARA) –

Archives and Mold Bibliography

New Brunswick Museum – Bibliography

North American Mycological Association – References for Teaching About Fungi

Regional Alliance for Preservation – Bibliography

Surviving mold – Mold-Sickness & Remediation Books

UNESCO – Prevention and treatment of mold in library collections with an emphasis on tropical climates: A RAMP study: Selected Bibliography

University of California, Riverside – Mycology

Bibliography (The Study of Fungi)

Western Wood Products Association – Mold Bibliography

World of Fungi – Bibliography

Google Books

Keyword search on the topic

DB BCIN

1. Save search term Mold and click BCIN below. Copy that word in the field 'Subject' and click 'Search'. 2. The same way do searches by terms mould, fungi, mildew. For displaying abstracts – select an article (articles) and press the button 'Viewing selected citations'

DB Conservation OnLine (CoOL)

Keyword search on the topic

Journal «Restaurator»

Keyword search on the topic

В начале раздела приводятся ссылки на библиографические указатели, после них – запросы для проведения самостоятельных поисков пользователями. Для поиска информации по зарубежным источникам предусмотрены четыре базы данных.

1) Conservation OnLine (CoOL). В ней содержится актуальная полнотекстовая информация по сохранности документов;

2) Bibliography of the Conservation Information Network (DB BCIN). Это самый полный сетевой библиографический ресурс по консервации и реставрации культурных ценностей, насчитывающий более 200 тыс. библиографических записей с рефератами;

3) БД журнала «Restaurator», наиболее авторитетного в рассматриваемой области. База данных содержит сведения о статьях, изданных в этом журнале, и обеспечивает коммерческий доступ к полным текстам статей;

4) База данных универсального назначения Google Books. При помощи этого сервиса можно предварительно просматривать книги, а в некоторых случаях – читать весь текст книги или загружать ее PDF-версию.

Библиографический поиск по русским источникам является менее полным и всесторонним, чем поиск по зарубежным источникам. В путеводителе предусмотрены ссылки на библиографический указатель «Консервация памятников культуры» (1958–2008) Российской национальной библиотеки, библиографический указатель по сохранности документов в БАН и полнотекстовая информация по консервации на сайте Art Conservation [14–18].

Другие регулярные рубрики путеводителя характеризуются следующим образом:

- класс **Dictionaries** включает специальные словари по каким-либо темам и/или словарные определения понятий. Ссылки на специальные словари в путеводителе имеются в том объеме, в каком словари представлены в Интернет. Информация из словарных изданий общего назначения представлена выборочно – определения понятий даются по наиболее авторитет-

ным энциклопедиям и словарям. Например, [1300 Pest Control – Pest Glossary](#), [A1 Pest Control – Pest Control Dictionary](#), [Absolute Serenity – Pest Control Glossary](#), [Accurate Pest Control – Davie Glossary](#), [Ask the Exterminator – Glossary of Pest Control](#), [Best Control 2 – An encyclopedia of Integrated Pest Management](#), [Beyond Pest Control – Pest Control Glossary](#), [Britannica Online Encyclopedia – Pest Control](#), [CAMEO: Conservation & Art Material Encyclopedia Online](#), [Colorado State University – Wildlife Damage Management Glossary](#), [Cornell University – Glossary of Biological Control](#), ... , [Oxford Dictionaries – Pest ...](#)

- класс **Databases** включает в себя информацию о базах данных по биологическим объектам и химическим веществам, применяемым как средства борьбы с вредителями библиотечных, архивных или музейных материалов. Например: [APVMA database – How to find the right insecticide](#), [Armed Forces Pest Management Board – Literature Database](#), [Cornell University – Pesticide Training Course Database](#), [Finding Dulcinea – Identifying Common Pests](#), [Goliathus – Online Insect Museum](#), [Harvard University – MCZ Entomology](#), [Inseclopedia.com – Databases](#), [International Database on Insect Disinfestation and Sterilization \(IDIDAS\) – Home page ...](#)

- в классе **Guidelines & Manuals** представлены ссылки на руководства и учебные материалы. Например: [Armed Forces Pest Management Board – Contingency Pest Management Guide \(Brochure\)](#), [BASF – Pest Management Guide: Plant Protection Products](#), [Beneficial Insects – Biological Insect Control Guidelines for Natural Pest Control](#), [Chartered Institute of Environmental Health – Pest Control Procedures Manual: Bedbugs](#), [Chicora Foundation – Managing Pests in Your Collections](#), [Collections Trust – Pest Management: A Practical Guide](#), [Control of Pest – Certified Pest Management Guide](#), [English Heritage – English Heritage guideline for insect pest management in English heritage historic properties](#), [Food and Agriculture Organization – Manuals & Guidelines](#), [Insect Pest Control ...](#)

- **Directories** включает в себя информацию о директориях, как [AFPMB Directory – Armed Forces Pest Management Board](#), [Alpha Ecological pest control company – Products](#), [Andre Pest Management Business Directory – Home page](#), [Bio-Integral Resource Center](#), [Business.com](#), [Canadian Pest Control Directory – Home page](#), [DMOZ – Home: Homeowners: Pest Control](#), [Inspection Angel – Pest Control & Building Inspection Services Directory](#), [Local Directories – Pest Control](#), [National Pest Management Association – Member Directory](#), [North Carolina Pest Directory – Termite and Pest Control Service Locator](#), [Open Directory – Science: Agriculture: Pests and Diseases](#), [Pest Control Alerts – IPM directory of least-toxic products](#), [Pest Control Directory – Pest Control Service](#), [Pest Control Canada – Canadian Pest Control Service professional directory ...](#)

- класс **Основные ссылки** по теме специального имени не имеет. Эти ссылки располагаются в коренном меню каждой рубрики за регулярными классами **Bibliographies**, **Dictionaries**, **Guidelines & Manuals**, **Databases**, **Directories** и разделителем «- - - - -».

В отличие от путеводителя по зарубежным источникам в русской части путеводителя выделяется только два класса, аналогичных вышеописанным: основная информация по какой-либо теме и класс «Библиографии и словари». Разная степень декомпозиции рубрик по русским и зарубежным источникам объясняется тем, что отечественной информации по консервации документов имеется значительно меньше, чем зарубежной. В силу этого нет оснований и материала для проведения столь же развернутой жанровой детализации отечественных тематических рубрик. Разной степенью декомпозиции объясняется также различие в количестве тематических рубрик. Русских рубрик в путеводителе имеется 73, а английских – 161.

4. ПОДГОТОВКА ЗАПРОСОВ ДЛЯ КОМПЛЕКТОВАНИЯ ПУТЕВОДИТЕЛЯ

Кроме описанных выше случаев декомпозиции понятий (в схеме предметной области и в перечнях тематических ссылок), декомпозиция регулярно используется также при подготовке информационных запросов, по которым ведется комплектование путеводителя. Для комплектования отбираются главным образом веб-страницы, где тема описана в заголовках страниц. Указанным образом обеспечивается высокая релевантность информационного поиска и уменьшение трудоемкости составления путеводителя. Средний объем информации по широким запросам при поиске в Google в рассматриваемой тематической области составляет 3–5 млн ссылок, а объем релевантной информации, содержащейся лишь в заголовках страниц, составляет около 100–150 ссылок. При описанном правиле комплектования путеводителя трудоемкость его составления сокращается до разумных пределов.

При поиске информации в Интернете запрос разбивается на подзапросы. Необходимость в указанной обработке определяют такие факторы:

- возможная синонимия информационных запросов;
- жанрово-видовая дифференциация сетевой информации и отражение ее при подготовке информационных запросов;
- необходимость уточнения запросов путем их привязки к тематической области консервации документов;
- излишне большой объем информации, стоящей за тематической рубрикой.

Синонимичные выражения понятий устанавливаются по тезаурусу по сохранности документов [19] или из получаемой из Сети информации. Для каждого из синонимов составляется информационный запрос, например, информации о бескислотном картоне соответствует следующий список синонимов (подзапросов):

Acid Free Cardboard
Acid Free Carton
Alkaline Cardboard
Alkaline Carton
Archival Cardboard
Archival Carton
Archival Quality Cardboard
Archival Quality Carton

Long-Lived Cardboard
Long-Lived Carton
Neutral Board
Neutral Carboard
Neutral Carton
Permanent Cardboard
Permanent Carton

К перечисленным терминам близко стоят словосочетания:

Archival Boxes
Archival Cases
Display Cases
Microclimate Boxes
Microclimate Containers
Museum Cases

По данным названиям также проводится поиск и отбор информации по соответствующей теме.

Декомпозицию той же темы по признаку 'жанрово-видовая характеристика информации' обеспечивают следующий блок подзапросов:

Bibliography Acid Free Cardboard
Bibliography Acid Free Carton
Bibliography Alkaline Cardboard
Bibliography Alkaline Carton
Bibliography Archival Cardboard

.....
Vocabulary (Dictionary) Acid Free Cardboard
Vocabulary (Dictionary) Acid Free Carton
Vocabulary (Dictionary) Alkaline Cardboard
Vocabulary (Dictionary) Alkaline Carton
Vocabulary (Dictionary) Archival Cardboard

.....
Guideline (Manual...) Acid Free Cardboard
Guideline (Manual...) Acid Free Carton
Guideline (Manual...) Alkaline Cardboard
Guideline (Manual...) Alkaline Carton
Guideline (Manual...) Archival Cardboard

.....
Directory Acid Free Cardboard
Directory Acid Free Carton
Directory Alkaline Cardboard
Directory Alkaline Carton
Directory Archival Cardboard

.....
Третья причина декомпозиции связана с уточнением содержания запроса путем привязки его к тематической области консервации документов. Уточнения требует содержание запросов большинства тематических рубрик, а именно содержание запросов, относящихся к классам:

Paper & Book Conservation
Environment Management
Pest Management
Disasters & Emergency

Вне контекстного окружения содержанию этих рубрик в Сети соответствует очень большое количество шумовой информации, относящейся не к сохранности документов, а к самой широкой экологической или сельскохозяйственной проблематике. Для привязки запросов к тематической области консервации документов в них «привносятся» необходимый контекст и используются ключевые слова со значением 'документы' и 'документохранилища'. Так, для

рубрики **Environment Management** такое уточнение ее содержания дает следующий ряд подзапросов:

Book «Climate Control»
Document «Climate Control»
Book «Environment Control»
Document «Environment Control»
Book «Environmental Monitoring»
Document «Environmental Monitoring»
Library «Climate Control»
Museum «Climate Control»
Archives «Climate Control»
Library «Environment Control»
Museum «Environment Control»
Archives «Environment Control»
Library «Storage Environment»
Museum «Storage Environment»
Archives «Storage Environment»

Содержание запросов, относящихся к классу **Pest Management**, уточняется с помощью тех же лексических средств и представлено в виде следующих подзапросов:

Biocide – Biocide Archives, Biocide Library, Biocide Museum, Biocides Paper, Disinfectant Archives, Disinfectant Library, Disinfectant Museum, Insecticide Archives, Insecticide Library, Insecticide Museum;
Disaster Preparedness – Disaster Preparedness Archives, Disaster Preparedness Library, Disaster Preparedness Museum, Emergency Recovery Archives, Emergency Recovery Library, Emergency Recovery Museum;
Disaster Recovery – Disaster Recovery Library, Disaster Recovery Archives, Disaster Recovery Museum, Emergency Recovery Library, Emergency Recovery Archives, Emergency Recovery Museum;
Leather Conservation – Leather Conservation Archives, Leather Conservation Library, Leather Conservation Museum, Leather Preservation Archives, Leather Preservation Library, Leather Preservation Museum, Parchment Conservation Archives, Parchment Conservation Library, Parchment Conservation Museum, Parchment Preservation Archives, Parchment Preservation Library, Parchment Preservation Museum, Vellum Conservation Archives, Vellum Conservation Library, Vellum Conservation Museum, Vellum Preservation Archives, Vellum Preservation Library, Vellum Preservation Museum.

Указанным образом ключевыми словами в запросах эксплицируется упоминавшийся ранее актуальный предмет, наиболее интересный для специалистов по консервации документов. Выше указывалось, что ему соответствует содержание **Documents [Books, Collections, Stacks]** – [Environment Management, Pest Management] или **Хранилища [Библиотеки, Архивы, Музеи]** – [Environment Management, Pest Management]. Данный предмет информации выражается с применением декомпозиции запросов.

Декомпозиция иногда применяется для уточнения отдельных понятий. Наиболее сложной и трудоемкой работой была декомпозиция темы запроса «Общие вопросы консервации». Содержание этой рубрики раскрывалось путем составления 63 подзапросов, описывающих, что именно мы имеем в виду, говоря об общих проблемах рассматриваемой тематической области и, в частности, следующих:

[Брошюры и книги, представленные в Интернете] – paper conservation brochure, paper preservation brochure,

book conservation brochure, book preservation brochure, paper conservation «table of content», paper preservation «table of content», book conservation «table of content», book preservation «table of content»;

[Обзоры литературы] – paper conservation review, paper preservation review, book conservation review, book preservation review, book conservation overview, book preservation overview, paper conservation overview, paper preservation overview, book conservation survey, book preservation survey, paper conservation survey, paper preservation survey;

[Общие вопросы консервации] – book conservation general, book preservation general, paper conservation general, paper preservation general;

[Достижения, успехи, развитие] – book conservation achievement, book preservation achievement, paper conservation achievement, paper preservation achievement, book conservation development, «book preservation» development, paper conservation development, paper preservation development;

[Основы] – book conservation fundamentals, «book preservation» fundamentals, paper conservation fundamentals, paper preservation fundamentals, book conservation basics, book preservation basics, paper conservation basics, paper preservation basics;

[Новый подход] – book conservation new approaches, book preservation new approaches, paper conservation new approaches, paper preservation new approaches;

[Перспективы, тенденции] – book conservation perspective, book preservation perspective, paper conservation perspective, paper preservation perspective, book conservation trend, book preservation trend, paper conservation trend, paper preservation trend;

[Стратегии, принципы] – book conservation strategy, book preservation strategy, paper conservation strategy, paper preservation strategy, book conservation principle, book preservation principle, paper conservation principle;

[Методологии] – book conservation methodology, book preservation methodology, paper conservation methodology, paper preservation methodology.

Рассмотрение декомпозиции в настоящей работе позволяет сделать такие выводы.

1. Использование декомпозиции запросов обеспечивает высокую полноту результатов информационного поиска (полноту информации в путеводителе), но отличается большой трудоемкостью. Технология поиска упрощается на стадии пополнения путеводителя. Из числа подзапросов, использовавшихся на этапе первоначального поиска, отбирается небольшое число наиболее результативных запросов, по которым проводится пополнение. Это существенно сокращает трудоемкость создания путеводителя.

2. Декомпозиции тематических рубрик (регулярная и иррегулярная) используются для минимизации их наполнения и сокращения числа документов, которые просматривают пользователи. Форматирование содержания рубрик на базе предметно-аспектной модели обеспечивает удобные средства навигации в путеводителе. Так как списки предметных рубрик отсортированы в алфавитном порядке названий организаций, проводящее поиск лицо может легко отобрать информацию по виду и статусу организаций: библиотеки, архивы, музеи – национальные, университетские, городские и пр. По названию организации можно также судить, например, о предположительной релевантности ссылки. Если какая-то организа-

ция является авторитетной в рассматриваемой области (как, например, Getty Conservation Institute, Library of Congress), то для пользователя, знающего предметную область, это должно означать и потенциальную ценность стоящей за рубрикой информации.

3. Декомпозиция рубрик фрагментирует ссылки на сравнительно небольшие разделы, обычно включающие около 100 тематических ссылок. Работа с такими объемами информации, по нашему мнению, не должна вызывать затруднений у пользователей.

СПИСОК ЛИТЕРАТУРЫ

1. Conservation OnLine. – URL: <http://cool.conservation-us.org>.
2. Handmade paper on the net. – URL: <http://handpaper.freyerweb.at/about.php?lang=en>.
3. Consortium of Academic and Research Libraries in Illinois. – URL: <http://www.carli.illinois.edu/preservation-resources>.
4. Preservation Readiness Webliography. – URL: <http://www.nedcc.org/free-resources/preservation-leaflets/6.-reformatting/6.8-digital-preservation-readiness-webliography>.
5. Preserving Access to Digital Information. – URL: <http://www.digitalpreservationeurope.eu/competence-centres/list/?id=127>.
6. Audio Preservation: A Beginner's Guide (PDF). – URL: <http://www.digitaltraditions.net/html/Resources/AudioPreservationGuide.pdf>.
7. Вчерашний Р.П., Космачева Н.В., Михайлова В.П. Об одном методе развития дескрипторного ИПЯ для патентного поиска // Электротехническая промышленность : реф. науч.-техн. сб. Сер. Общеотраслевые вопросы. – 1977. – № 12 (487). – С. 16–19.
8. Вчерашний Р.П. Системы и средства патентного поиска. – М. : ЦНИИПИ, 1977. – 96 с.
9. Розенсон Г.С., Вчерашний Р.П., Сологуб Г.Н. Использование типовых блок-схем декомпозиции объектов техники для индексирования тематики НИОКР по МКИ // Проблемы развития государственной системы патентной информации : тез. докл. и сообщ. – М., 1977. – С. 184–187.
10. Поиск в Интернете. Все студенту. – URL: <http://everything-for-students.ru/?p=306>.
11. Как пользоваться поисковиками. – URL: <http://www.nejaden.ru/google/>.
12. Стратегия и методика профессионального информационного поиска. – URL: <http://textbook.vadimsteva nov.ru/chapter2/glava2-3.html>.
13. Пименов Е.Н. Семантико-синтаксическая модель индексирования документальной информации // Структурная и прикладная лингвистика. – СПбГУ, 2004. – Вып. 6. – С. 278–294.
14. Пименов Е.Н. Особенности языка индексирования в БД по сохранности документов // Материалы IV научно-практического семинара «Электронные ресурсы библиотек». – СПб., 2008. – С. 89–94.

15. Пименов Е.Н. Предметно-аспектный подход к индексированию информации: актуальный и нулевой предмет индексирования // Научно-техническая информация. Сер. 2. – 2001. – № 7. – С. 18–25.
16. Библиографический указатель по консервации памятников культуры Российской национальной библиотеки. – URL: <http://www.nlr.ru/fdcc/publ/>.
17. Библиографическая база данных по сохранности документов Библиотеки Российской академии наук. – URL: http://www.ras.ru/e_resours/Consev_restavr/.
18. Полнотекстовая информация по консервации на сайте Art Conservation. – URL: <http://art-con.ru/>.
19. Тезаурус информационно-поисковый по сохранности документов / сост.: Е.Н. Пименов, Л.Г. Левашова, В.Б. Никитин. Б-ка РАН. – СПб.: БАН, 2006. – 295 с.

Материал поступил в редакцию 01.06.15.

Сведения об авторе

ПИМЕНОВ Евгений Николаевич – кандидат филологических наук, старший научный сотрудник Отдела информатики и автоматизации Библиотеки Российской академии наук, Санкт-Петербург
e-mail: pen48@list.ru

СПИСОК ОПЕЧАТОК

к статье В.К. Финна «Обнаружение эмпирических закономерностей в последовательностях баз фактов посредством ДСМ-рассуждений», опубликованной в сборнике НТИ, сер. 2, № 8. - С. 1-29.

Стр.	Колонка	Строка	Следует читать
3	правая	25 снизу	... данную посредством рекуррентного соотношения
4	правая	19 снизу	$M_{ab,n}^{\sigma}(V, W), M_{ad_0,n}^{\sigma}(V, W), M_{ad_2,n}^{\sigma}(V, W)$, где $\sigma \in \{+, -\}$
4	правая	17,18 снизу	сходства с запретом на контрпримеры, различия, сходства-различия.
4	правая	1 снизу	$\forall V \forall W M_{x_1,n}^{+}(V, W) \rightarrow M_{x_2,n}^{+}(V, W)$, где « $\rightarrow$ » -
7	правая	24 сверху	$AKP^{(+)} \forall X \forall Y \exists k \exists V_1 \dots \exists V_k \exists W_1 \dots \exists W_k (J_{(1,0)}(X \Rightarrow_1 Y) \rightarrow$
7	правая	28 снизу	$AKP_{*}^{+} \forall X \forall Y \exists V (J_{(1,0)}(X \Rightarrow_1 Y) \rightarrow \exists n (J_{(1,n)}(V \Rightarrow_2 Y) \&$
7	правая	22 снизу	формулой $J_{(1,0)}(X \Rightarrow_1 Y)$, существует гипотеза ДСМ-
8	левая	10 снизу	ставимы посредством $J_{(v,i)}(V \Rightarrow_1 W)$ и $J_{(t,i)}(V \Rightarrow_1 W)$
10	левая	20, 21 снизу	$\mu = \begin{cases} 1, & \text{если } \sigma = - \\ -1, & \text{если } \sigma = + \end{cases}$
12	левая	22, 23 снизу	$f^{+}(p, q) = \frac{ \Delta^{+}(p) \cap (\Delta^{-}(q) \cup \Delta^{0}(q)) }{ \Delta^{+}(p) }$
12	левая	21, 20 снизу	$f^{-}(p, q) = \frac{ \Delta^{-}(p) \cap (\Delta^{+}(q) \cup \Delta^{0}(q)) }{ \Delta^{-}(p) }$
12	левая	19, 18 снизу	$f^{0}(p, q) = \frac{ \Delta^{0}(p) \cap (\Delta^{+}(q) \cup \Delta^{-}(q)) }{ \Delta^{0}(p) }$
14	левая	6 снизу	Утверждения (15-2) и (16-2) выражают устойчи-
14	правая	9 сверху	$\bar{R}(p, q)$
15	левая	4 сверху	$\& \exists q \exists l (J_{(1,l)}(V \Rightarrow_2 W) \in \Delta^{+}(p))$;
15	левая	7 сверху	$\& \exists q \exists l (J_{(1,l)}(V \Rightarrow_1 W) \in \Omega^{+}(p))$.
16	сноска	2 снизу	¹ В данной статье T_a есть две булевы алгебры B_1 и B_2
16	сноска	1 снизу	² Аксиомы из $T_a \cup T_{pr}$ представлены в §4
16	правая	2 сверху	+ п.п.в.-2) n_p , где n_p – номер шага стабилизации пре-
20	правая	8 сверху	Для строк Табл. 3 и 4 $\pi_i^{(2)}$ и $\pi_i^{(1)}$ определим, соот-
20	правая	15 сверху	$B_1^{+}(i) \Leftrightarrow \{\sigma_i(p) (\sigma_i(p) = 1) \& (1 \leq i \leq l) \& (0 \leq p \leq s)\}$,
22	правая	2 снизу	Если заданы $Des_{j,i}^{\sigma}$, ξ_j^{σ} , равенства $\xi_j^{\sigma} = 1$ или не-
23	правая	14 снизу	$L_{1,2} = L_{1,2}^{+} \cup L_{1,2}^{-}$ и $L_{3,4} = L_{3,4}^{+} \cup L_{3,4}^{-}$; аналогично
24	правая	13 сверху	(т.е., $\mu(s)$ есть $\mu_{x,y}(s)$), пусть далее, для $Str_{x,y} l_0, a, b, c$
24	правая	25 сверху	вом Ω_t , где $ \Omega_t = l_0$, $0 \leq l_0 \leq m_0$, а $m_0 = \Omega_{0,0}^{\tau} \geq \Omega_t = l_0$.
24	правая	32 сверху	В Табл. 6 $\sigma_j^{(h)} \in \{0, 1\}$, $h = 1, \dots, d, j = 1, 2, 3, 4, 5, 6$;
25	левая	10 снизу	выводов – индукции (п.п.в.-1) и аналогии (п.п.в.-2) – яв-