Концепция и средства информационной среды поддержки когнитивных процессов

© О.Л. Голицына, Н.В. Максимов

Национальный исследовательский ядерный университет «МИФИ», г. Москва

OLGolitsina@yandex.ru, NV-Maks@yandex.ru

Аннотация

С позиций системного анализа обсуждается архитектура информационной среды, ориентированной на поддержку процессов синтеза знаний, которая, помимо хранения, поиска и систематизации информации, должна иметь средства динамического построения и использования компонентов лингвистического обеспечения, а также средства анализа как результатов поиска, так и тенденций научных направлений.

1 Введение

Для инновационного развития важнейшее значение имеет задача эффективной информационной поддержки исследовательских работ и процессов подготовки научно-педагогических кадров. Сложность решения этой задачи обусловлена особенностями информационных пространств образования и науки, наличием большого разнообразия документов и массивов, хранящихся и обрабатываемых в различного вида информационных системах, библиотеках и архивах. Заметное влияние стали оказывать экономические и правовые факторы, делающие недоступными технически доступные в сети и потому, вроде бы открытые, научные публикации. Отметим, что такое состояние дел породило тенденцию открытой авторской публикации научной информации и формирования социальных сетей (см., например, [1 – 3]).

Другим фактором, существенно затрудняющим эффективное использование информации, является свойство рассеяния, характерное для всех этапов жизненного цикла генерации/использования знаний.

Объекты исследований распределены как по всем этапам жизненного цикла, так и среди различных субъектов – индивидуальных или коллективных авторов; описания объектов, процессов и результатов исследований распределены в лексическом пространстве (для описания одного и того же объекта разные авторы могут, причем одинаково эффективно, использовать разную лексику); сведения о публикациях распределены среди разных справочных изданий, баз данных и сайтов, а экземпляры публикаций – среди разных хранилищ; форма и способ представления информации распределены среди различных международных, национальных и фирменных стандартов на наполнение и формат документов; функции поиска и обработки информации распределены среди многих реализаций поисковых систем, а их интерфейсные представления – среди различных метафор и визуальных компонентов, используемых разработчиками.

Наконец, возможности человека по переработке информации тоже сильно ограничены, и поэтому более или менее большой поток разнообразных сообщений, если они не упорядочены явно в соответствии с некоторой схемой (темой), не будет полноценно связан с наличным знанием. Показательным примером является поиск в интернет-машинах. Формируемая ими выдача, всегда большая и неадекватная[1] (если рассматривать её как целостный ответ на конкретный запрос) и потому избыточная для восприятия, упорядочена обычно в соответствии со «знаниями» поисковой машины, а не с семантикой потребности, что в итоге уводит субъекта от систематического формирования понятийной структуры и порождает «клиповое сознание».

Традиционно в задачах информационного обеспечения науки и управления для организации информационных потоков используется лингвистическое обеспечение (ЛО) – классификаторы, рубрикаторы и тезаурусы, являющиеся метаинформационными компонентами информационно-поисковых систем (ИПС). Именно ЛО позволяет не только более или менее эффективно и единообразно идентифицировать (определить место) содержание того или иного информационного сообщения, но и отра-

[1] По существу, широко используемые поисковые интернет-машины в случае научного поиска, выполняемого информационно непрофессиональным пользователем (не знающим законы рассеяния и инфраструктуру информационной отрасли), этого пользователя *псевдоинформируют*, подтверждая «обстоятельность» и «значимость» найденного подавляющими все сомнения количественными показателями. Экспериментальные данные, в том числе и полученные авторами, показывают, что найденные объемы релевантной информации составляют доли процента от того, что может быть найдено в проблемно-ориентированных профессиональных БД

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

жает системность организации науки, фиксируя общее и устойчивое представление о составе и взаимосвязях отдельных разделов и направлений исследований.

Вместе с тем, создание и поддержка такого рода метаинформационных средств ныне сталкивается с существенными затруднениями. Информационная сфера, где этой (довольно ресурсоемкой!) деятельностью в эпоху индустриального общества целенаправленно занимались международные и национальные службы с привлечением экспертов всех отраслей знаний, сейчас уже не имеет возможностей полно и оперативно реагировать на развитие науки. Характерно, что и в информационных задачах пользователи перешли на режим «самообслуживания», что не улучшило обратную связь, позволяющую поддерживать ЛО в адекватном состоянии.

В этом смысле задача создания и внедрения распределенных технологий «естественного» формирования информационных и метаинформационных компонентов является актуальной.

2 Концептуальные основы

Для определения подходов к созданию распределенных технологий формирования и использования информационных и метаинформационных компонентов будем рассматривать обобщенную систему воспроизводства и преобразования знаний, в которой автоматизированная ИС составляет часть основной деятельности – создания нового знания.

Примем согласно [4], что в основе модели синтеза знаний как самоорганизующегося процесса лежит структурная особенность системы – возможность ее разложения на относительно независимые подсистемы. Таким образом, сложная система может быть описана при помощи набора относительно независимых аспектных представлений (контекстов, построенных над «сеткой» базовых понятий и отношений). Каждое такое описание дает лишь частичное знание о системе в целом, но полное по отношению к данному аспекту. Соответственно, объединение и согласование различных контекстов позволяет построить целостное представление о системе. Существенно, что в процессе декомпозиции не только выделяются составляющие, но и формируется *схема декомпозиции* – система характеристических признаков, в соответствии с которой и проводится декомпозиция.

Эта методика по своей сути является реализацией системного подхода, позволяющего, с одной стороны, представить объект как множество[2] однородных (типизированных), связанных некоторыми отношениями элементов, в совокупности образующих единство, а с другой – представить эту совокупность в виде классификации, что, в свою очередь, дает возможность выделять в явной форме новые характеристические признаки, определять способы выделения подсистем и на основе свойств соответствия и симметрии обнаруживать связи с другими системами классификации [5]. Именно такой подход методологически связывает относительно самостоятельные и в то же время взаимообусловливающие объекты основной и информационной деятельности – документы и классификации[3].

ИПС в обобщенной человеко-машинной системе основной/информационной деятельности играет замещающую роль, и поэтому поиск потенциально полезной информации с точки зрения теории систем может рассматриваться как процесс построения новой *системы* знаний, где «технологической» основой и своеобразным *методом генерации* информации является комбинаторное сочетание, а ИПС выполняет роль «перемешивающего слоя», формируя неравноценные комбинации информационных компонентов (выборки документов и терминов) и стимулируя, тем самым, ускорение возникновения неравновесного состояния [6].

Особенностью распределения функций между человеком и автоматизированной информационной системой состоит том, что основные определяющие функции – выбор цели, определение критерия полезности, оценка и принятие решения, а, главное, генерация новой информации, осуществляются человеком. Только человек знает, что ищет. Только он может образовывать или выделять как *проблемные ситуации*, так и *полезные* ассоциативные связи между различными информационными объектами и выводить на их основе новые свойства, т. е. поисковые механизмы ИПС готовят альтернативы, а средства систематизации и протоколирования задают (точнее, фиксируют) направления развития, технологически позволяя пользователю выбирать (а не генерировать) «предпочтительные».

С другой стороны, и человек знает больше, чем «публикует» (невербализованная составляющая). Именно здесь итерационное взаимодействие человека и информационной системы позволит «вытянуть» информацию об объекте исследования не только из информационной системы, но и из сознания человека. Система, фиксируя траекторию поисков и информационные образы (ею сгенерированные, но выбранные человеком), позволяет не только в любой момент вернуться к любому информационному объекту и пойти по другой траектории, но и вербализует неявные знания человека.

[2] Совокупность таких представлений с точки зрения общей теории систем может быть определена как система $S_{i,} =< M_i, A_i, R_{i,} Z_i >$, $i=1,\ldots,n,$ где i – аспект, который отвечает своему закону композиции Z_i, связывающему множество элементов M_i, определенных на множестве характеристических признаков A_i и связанных отношениями R_i

[3] В этом смысле развивающееся познание можно представить двойной спиралью, в которой эволюция фактов (гипотез, методов, результатов), представленная документами, синхронизирована с эволюцией системной точки зрения (парадигмами, организацией знаний и науки), представленных классификациями и понятийными системами. Именно это обеспечивает при фрагментарной природе процесса познания целостность и устойчивость его развития

Технологически этот процесс поддерживается средствами *когнитивного рубрикатора* [7] – динамически создаваемой пользователем иерархически[4] организованной структуры, которая будет интенсионально (через систему классификационных признаков) и экстенсионально (через подборки документов, фрагментов понятийных и терминологических систем) представлять *индивидуальные знания,* соотнесенные с общепринятыми представлениями предметной области (ПрО). Интегральность такого представления достигается за счет того, что оно (1) реализуется объектами как уровня ресурсов – упорядоченными подборками документов, ссылками на ассоциированные ресурсы и т. д., так и уровня терминологии – запросами, словниками, фрагментами рубрикаторов и тезаурусов, используемых в данной ПрО, и (2) явно фиксирует общность и различия представлений ПрО, характерные для конкретной проблемной ситуации, как с точки зрения полноты и специфичности представления объекта информационной потребности различными терминологическими системами, так и с точки зрения характера её представления в различных ресурсах, в том числе отражаемого динамикой развития ПрО (временными рядами потоков публикаций и лексики для различных составляющих и аспектов предметной области).

Таким образом, *рабочее пространство пользователя WS* может быть описано тройкой $WS = \langle IR_w, IR_U, F_{WS} \rangle$,

где

IR_W – доступные локальные и внешние (мировые) информационные ресурсы;

IR_U – информационные ресурсы, создаваемые пользователем;

F_{WS} – классификационная схема, организующая рабочее пространство и отражающая личный взгляд пользователя на ПрО, реализованная в форме когнитивного рубрикатора, фиксирующего знания пользователя о ПрО на понятийном, документальном и лексическом уровнях.

Исходя из общего принципа индексирования каждого документа в рабочем пространстве, для описания совокупности IR_W и IR_U можно использовать линейную модель представления документов терминами универсального словаря.

Рубрикатор имеет иерархическую организацию и может быть структурно представлен в виде ориентированного дерева – ациклического орграфа G (в котором только одна вершина – корень дерева – имеет нулевую степенью захода).

В свою очередь рубрикатор F_{WS} описывается двумя матрицами: $F_{WS} = \langle R, M \rangle$, где

R – бинарная матрица «рубрика-документ»;

M – матрица смежности дерева G.

Для когнитивного рубрикатора определены операции преобразования графа, а также предложены дистрибутивно-статистические оценки сбалансированности и логической непротиворечивости структуры представления ПрО.

3 Практическая реализация интегрального подхода

В качестве программной основы использовалась документальная информационно-аналитическая система xIRBIS [8], интегрированная с системой лингвистического анализа АОТ и системой машинного перевода RETRANS. В составе системы выделяются пять подсистем, обеспечивающих как традиционные функции создания ИР и информационного поиска, так и поддержку понятийно-терминологических систем.

Подсистема информационного поиска в распределенных ресурсах помимо классических механизмов поиска по четким и нечетким критериям и с реформулированием запроса по обратной связи обеспечивает переадресацию и адаптацию запроса для проведения поиска в других ресурсах с учетом их особенностей, в том числе синтаксиса поисковых языков[5].

Подсистема логико-семантического анализа осуществляет в автоматизированном интерактивном режиме построение понятийного образа научного документа[6]. Эти процедуры, по существу, реализуют принцип дополнительности: представление знания в виде извлекаемых из текста ключевых слов и отношений ИПС осуществляет с точки зрения «устоявшейся» системы понятий (статистической значимости), а человек, внося изменения и дополнения

[4] Накопленные знания, как и процесс познания в информационных и когнитивных задачах, представляются обычно иерархическими структурами ЛО, выполняющими роль своеобразной системы координат ПрО. Будучи достаточно простыми, они конструктивно соединяют «логику» и «физику» когнитивных процессов: индукцию/дедукцию, декомпозицию/синтез. Кроме того, иерархия хорошо соответствует механике процесса познания: углубление знаний осуществляется по схеме специализации обычно путем деления текущего целого по крайне мере на две части в соответствии со значениями выбранного признака деления. Однако такое деление, рассматриваемое с точки зрения развивающейся ПрО, будет корректно только для фиксированного, уже состоявшегося знания, а не того, которое, возможно, будет, т. е., выбирая ПрО (производя необратимую «редукцию» возможных состояний информации) и, тем самым, определяя «главные направления», мы фактически формируем еще и «мнимую», остающуюся вне процесса познания, область. Это означает, что производится не целочисленное, а «дробное» деление предметной области и, соответственно, структура её представления имеет не иерархическую, а, скорее, фрактальную природу

[5] Основные положения этой подсистемы рассмотрены в представленном на RCDL'2010 докладе Окропишина А.Е. «Об одном подходе к организации документального поиска в распределенных гетерогенных информационных ресурсах»

[6] Основные положения рассмотрены в представленном на RCDL'2010 докладе Окропишиной О.В. «Технология автоматизированного формирования понятийной структуры научного контента»

в граф, построенный системой, фиксирует отличия, характеризующие новизну и специфику по отношению к общепринятому представлению.

Подсистема статистического анализа документальных потоков и лексики обеспечивает формирование распределений различных информационных срезов, а также построение и анализ временных рядов профилированных потоков документов и лексики.

Подсистема анализа и ведения лингвистического обеспечения ориентирована на поддержку пользовательского лексического пространства предметной области и обеспечивает построение и ведение иерархических словарных структур, которые могут быть использованы в качестве мини-тезаурусов, индивидуализирующих ПрО, а также для автоматической классификации документов.

Особенность технологических решений информационно-аналитической системы xIRBIS в том, что полнота отбора информации обеспечивается не только использованием различных информационных ресурсов. Поисковые технологии учитывают двойственность природы форм и способов представления запроса: при общем стремлении к «завершенной» вербальной форме выражения запроса в силу неопределенности, присущей реальной потребности, часть или даже весь запрос может быть представлен в форме отдельных документов или их кластеров. Механизмы поиска также построены по принципу дополнительности четких и нечетких моделей. При этом нечеткие механизмы, реализующие кластеризацию документов на основе обучения на примерах, дополняются технологиями динамического реформулирования запроса по обратной связи по релевантности. Это позволяет, с одной стороны, выделять статистически значимые подмножества, а с другой – выявлять документы пограничные, статистически не значимые, но, возможно, обладающие существенным признаком новизны. Анализ документов, получаемых по запросам, позволяет на следующих этапах не только оценить возможность целевого практического использования их содержания, но и обогатить запрос, а также расширить терминологию предметной области. В свою очередь, систематизация терминологии и анализ потоков информации позволяет динамически строить понятийную модель ПрО, являющуюся не только необходимым элементом познавательного процесса, но и одной из форм представления, сохранения и распространения знаний.

4 Заключение

Вышеприведенное позволяет сделать вывод, что, по существу, информационной системой может называться только *совокупная* система, объединяющая генераторов (в роли которых выступает обычно человек) и поставщиков-посредников (в роли которых выступает автоматизированная система) информации. Без средств систематизации, позволяющих не только упорядочить массив, но и обобщенно отразить его содержание в системе наук, языка запросов, позволяющего «позиционировать» индивидуальность точки зрения пользователя, и, наконец, человека, который, комбинируя и систематизируя получаемые данные и наличные знания, синтезирует новое знание, любой информационный фонд – это только хранилище данных.

Система поиска информации в документальных ресурсах по существу трансформируется в систему управления навигацией в среде информационных компонентов, отражающих знания на всех уровнях: лексическом, понятийном и документальном. Такая «*гиперинформационная*» система обеспечивает переходы не только между объектами одного уровня (текстами документов, справочников и т. д.), но и объектами, относящимися к разным уровням, в том числе виртуальным (динамически создаваемым статистическим выборкам и временным рядам): например, от слова – к онтологии или документам, от множества документов – к словнику ПрО, от точки временного ряда – к соответствующим документам.

Рассмотренные решения представляют собой не только инструменты распределенного поиска информации, но и технологию распределенной экспертной оценки научной работы. Автор, формулируя основные положения работы, выделяет не только основные общепринятые понятия, но и новые; эксперты формируют не только оценку работы в целом, но также оценивают её уровень и место в предметной области.

В целом такая технология за счет системного использования лингвистических средств и механизмов поиска позволит *пользователю* (ученому или разработчику, а не специализированной информационной службе) создавать проблемно-ориентированный информационный ресурс, им самим *формируемый* и *систематизируемый*, который может включать помимо подборок документов также и метаинформацию, например, словари специальной терминологии, «индивидуальные» классификаторы предметных областей, описания ресурсов, библиометрические и другие срезы и т. д.

Рассмотренные процессы в целом представляют интегральную распределенную технологию[7] формирования не только информации об отдельной научной работе, но также и аналитической информации о научных направлениях, экспертах и научных коллективах, а косвенно, и информации для актуализации метаинформации, систематизирующей научные направления (рубрикаторов научных направлений, классификаторов специальностей и т. д.). Это

[7] Рассматриваемая технология была реализована в рамках Федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2012», проект «Разработка и внедрение информационно-аналитической системы регистрации, учёта, обработки и хранения отчётных документов по НИОКР с целью проведения мониторинга состояния и основных тенденций и направлений развития научных исследований и разработок» [9]

означает, что в информационной среде наряду с двумя «традиционными» потоками – генерации/публи-кации информации и обслуживания запросов – в явной форме реализуется поток вариантных (с разной степенью общности/отличий) представлений предметных областей, семантически и технологически связывающий две, также традиционные, но достаточно изолированно существующие формы представления ПрО – лингвистическое обеспечение ИС, централизованно создаваемое и поддерживаемое экспертными комиссиями, и понятийно-терминологические системы отдельных ПрО, создаваемые авторами.

Именно технология интеграции распределенных процессов основной и информационной деятельности в общей среде, с одной стороны, и использование некоторых форм представления объектов и результатов ОД в качестве информационных ресурсов – с другой, позволяют говорить о переходе от отдельных специализированных форм информационного обслуживания к интегральным (по функциям) и интегрированным (по процессам основной и информационной деятельности) системам. В итоге это создаст технологические и методологические условия для формирования не только информационной среды генерации, обработки и хранения знаний, обеспечивающей преемственность и развитие информационной поддержки процесса познания на всех этапах его жизненного цикла, но также и «распределенного» экспертного сообщества.

Литература

[1] Горбунов-Посадов М.М. Интернет-активность как обязанность ученого. – М.: ИПМ им. М.В. Келдыша, 2007. – http://library.keldysh.ru/prep_vw.asp?pid=2858.

[2] Паринов С.И. e-Science – онлайновое будущее науки. //Информационные технологии. – 2007. – №9.

[3] Паринов С.И., Когаловский М.Р. Технология поддержки электронных научных публикаций как «живых» документов. // Труды 11й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», RCDL'2009, Петрозаводск, Россия, 2009. – С. 53-58.

[4] Яблонский А.И. Модели и методы исследования науки. Серия: Философы России XX века. – М.: Едиториал УРСС, 2001.

[5] Урманцев Ю.А. Общая теория систем: Состояние, приложения и перспективы развития// Сб. «Система, Симметрия, Гармония». – М.: Мысль, 1988. – С. 38-124.

[6] Чернавский Д.С. Синергетика и информация. – М.: Едиториал УРСС, 2004.

[7] Голицына О.Л., Максимов Н.В. Об архитектуре и программно-информационных средствах поддержки когнитивных процессов // Информационные технологии в образовании. XIX межд. конф.-выставка: Сборник трудов, Ч. II. – М.: МИФИ, 2009. – С. 22-25.

[8] Максимов Н.В., Васина Е.Н., Голицына О.Л. и др. Документальная информационно-аналитическая система xIRBIS: программа для ЭВМ// Свидетельство о гос. регистрации №2008611511 от 25.03.2008.

[9] Максимов Н.В., Павлов Л.П., Строгонов В.И. Интегральный подход к информационно-аналитическому обеспечению науки и образования // Системы управления и информационные технологии. – 2009, № 4.1. – С. 165-169.

Conception and means of information's environment support of cognitive processes

O.L. Golitsina, N.V. Maksimov

The architecture of information environment, that oriented toward support of knowledge synthesis, are considered in contexts of system analysis. The information environment, besides search and systematization of information, must have means of dynamic synthesizing of lingware, as well as the analysis facilities for search data and research trends.