КОНТЕКСТНЫЙ АНАЛИЗ СОБЫТИЙ И СИНТЕЗ СТРУКТУРЫ МЕДИЦИНСКИХ ЗНАНИЙ

Я.И. Гулиев, В.Л. Малых, С.Г. Юрченко
Институт программных систем РАН, Переславль-Залесский, Российская Федерация

Целью работы является постановка задачи синтеза структуры медицинских знаний на основе стандартов обмена данными в медицине, а также на основе накопленных в МИС знаний, формализованных в виде объектов БД и структурных моделей клинических документов. Знания, используемые для синтеза структуры, предварительно анализируются и формализуются в виде событий – кортежей атомарных фактов. На атомарных фактах для передачи семантики контекстных отношений вводится отношение нестрогого порядка. Предлагается алгоритм синтеза нормализованной в некотором смысле структуры знаний. Полученные результаты могут быть полезны для разработки общей структуры медицинской карты, структуры мобильных медицинских карт. *Представленные исследования проводились в рамках гранта РФФИ № 07-01-00719-а.*

Введение

С задачей концептуализации и формализации медицинских знаний сталкиваются практически все разработчики медицинских информационных систем (МИС). В результате решения подобных задач появляются тезаурусы для описания используемых в медицине понятий, появляются стандарты обмена медицинской информацией, например, широко известный Health Level 7 (HL7). В рамках МИС возникают различные информационные структуры для хранения данных о событиях лечебно-диагностического процесса (ЛДП), возникают различные структурные модели медицинских документов [1], проектируются различные объекты базы данных (БД). Несмотря на многочисленные усилия, процесс концептуализации и формализации медицинских знаний далек от своего завершения. По мнению авторов работы [2], процесс концептуализации знания должен быть непосредственно встроен в МИС. В последнее время в медицинской информатике возрос интерес к построению достаточно общей структуры «всего» накопленного знания. Например, разработка структуры полной, включающей «все» данные, единой медицинской карты (ЕМК). Или, разработка полного, включающего все виды данных, стандарта обмена медицинской информацией. Для решения подобных задач необходимо выработать общий подход к структуризации медицинских данных. Изложение возможного подхода к концептуализации и формализации медицинских знаний дается ниже, а также намечаются пути и перспективы его дальнейшего развития.

1. Базис концептуализации

В основе базиса лежат понятия. На основе понятия определим атомарный факт (АФ) как пару понятий.

$$a = (a_1, a_2), \quad a_i \in A. \tag{1}$$

Где под множеством A понимается множество конечных последовательностей символов некоторого алфавита (тексты). Первое в паре понятие является общим абстрактным понятием: ‘пациент’, ‘диагноз’, ‘группа крови’. Второе понятие является конкретизацией первого абстрактного понятия: ‘Иванов’, ‘ОРЗ’, ‘IV’.

В качестве базиса представления знаний рассматриваются события S, которые конструируются из атомарных фактов и являются конечными последовательностями атомарных фактов.

$$S = \{a^i\}, \quad i \in N, \quad a \in A^2. \tag{2}$$

В нашем представлении, весь лечебно-диагностический процесс может быть представлен однородно в виде потока (множества) событий. Клинические документы, отражающие различные эпизоды ЛДП, также могут быть представлены в виде событий. Для усиления се-

мантической силы базиса концептуализации введем в рассмотрение понятие контекста атомарного факта a в событии S, считая далее всюду, что нумерация атомарных фактов в событии начинается с 1. Локальным контекстом атомарного факта $a^p \in S$ является событие K, являющееся подмножеством события S, и определяемое как

$$K = \{a^i\}, \quad i = \overline{1, p-1}, \quad K \subset S. \tag{3}$$

Содержательно локальный контекст – это другие атомарные факты, которые позволяют расширить семантическую интерпретацию данного атомарного факта. Например, атомарный факт {('температура', '38,5 С')} вне контекста события {('пациент', 'некто'), ('дата и время', '03.10.2008 13:00'), ('температура', '38,5 С')} мало что значит сам по себе. В нашем определении предполагается, что все атомарные факты предшествующие данному факту в событии являются содержательным контекстом факта. Но одновременно с этим можно сделать вывод, что неявно предполагается, что последующие атомарные факты «не требуются» для семантической интерпретации данного факта! Необходимо уточнить семантику нашего определения, чтобы не приписывать ему излишнюю силу. Очевидно, что интерпретация данных может потребовать самого широкого нелокального контекста, тем более, когда речь идет о медицине. Лечащий врач совместно интерпретирует множество событий ЛДП для правильной постановки диагноза, для выбора правильной тактики лечения. Наше определение контекста более узкое, «локальное». Достаточно понимать локальный контекст атомарного факта, как «достаточное» дополнительное множество других атомарных фактов, которые уже позволяют дать данному факту некоторую «прагматически ценную» семантическую интерпретацию.

Дадим еще одно, чуть более широкое, определение локального контекста атомарного факта. Предварительно заметим, что в приведенном выше примере для интерпретации атомарного факта ('температура', '38,5° С') нам не важен порядок двух других предшествующих ему атомарных фактов. Можно легко представить себе случай, когда несколько атомарных фактов могут быть интерпретированы только совместно. Например, для двух лабораторных тестов норма определяется совместно, как некоторая область на плоскости. Формально это означает, что атомарный факт b входит в контекст атомарного факта a, и одновременно атомарный факт a входит в контекст атомарного факта b. Чтобы учесть такие случаи дадим другое определение локального контекста атомарного факта a в событии S. Для этого введем на атомарных фактах, входящих в S, отношение нестрогого порядка $\leq$. Локальным контекстом атомарного факта $a^p \in S$ является событие K, являющееся подмножеством события S, и определяемое как

$$K = \{a^i\}, \quad a^i \leq a^p, i \neq p, \quad K \subset S. \tag{4}$$

Будем в дальнейшем считать, что в событии S в последовательности $\{a^i\}$ атомарные факты упорядочены, т.е. $a^i \leq a^{i+1}, i = \overline{1, n-1}$, где n число атомарных фактов в событии S.

Отметим, что в нашем подходе предполагается последовательное линейное «повествовательное» изложение фактов, когда все, или почти все (с учетом нестрого отношения порядка) необходимое для понимания очередного факта уже было сказано. Такой стиль изложения вообще свойственен логическому научному подходу, хотя он и не укладывается в гипертекстовое компьютерное представление информации. Отметим, что и сам поток событий ЛДП также упорядочен в первую очередь благодаря наличию у событий темпоральных свойств: времени начала, длительности протекания, времени завершения, и существованию естественного отношения порядка относительно временной оси. Поэтому в контекстных отношениях атомарных фактов могут проявляться и темпоральные отношения, одни факты во времени будут предшествовать другим.

Над событиями можно было бы определить, как над множествами, операции декомпозиции и объединения, позволив конструктивно выделять из события «подсобытия» и конструировать из событий охватывающие объединенные события. Но поскольку при этом не должна нарушаться семантика контекстных отношений, эти операции не могут выполняться исключительно формально без анализа результата на семантическую корректность.

2. Методология концептуализация

Перейдем к рассмотрению источников атомарных фактов и событий. Для разработчиков МИС основным источником атомарных фактов и событий являются клинические медицинские документы. Документы могут иметь явную структуру в виде оглавления, или разделов с подразделами. Клинические документы зачастую следуют линейному повествовательному стилю изложения фактов, что с точки зрения нашего подхода облегчает задачу их концептуализации. Другим важным источником уже концептуализированных и формализованных данных является БД МИС. Обычным часто наблюдаемым недостатком, допущенным при проектировании БД, является отсутствие в ней словаря (тезауруса) понятий, связывающего явно таблицы и атрибуты таблиц с понятиями предметной области. Еще одним препятствием на пути автоматической разборки словаря объектов БД, и превращения этих объектов в модели событий, является отсутствие на объектах БД требующихся нам контекстных отношений. Между таблицами реляционной БД могут существовать отношения, но при отражении таблиц в события ссылки на другие таблицы нельзя обрабатывать формально без учета семантики. Например, реляционная модель международного классификатора болезней (МКБ-10) может представлять собой таблицу со ссылкой на себя же для отражения иерархической структуры классификатора. Формально с помощью этой реляционной структуры можно передать иерархию любой глубины, но в МКБ-10 эта иерархия заведомо ограничена, а сами уровни иерархии имеют различную семантику: класс, рубрика, подрубрика. В любом случае процесс концептуализации и формализации медицинского знания не может выполняться без участия аналитиков и специалистов предметной области. Возложить эту задачу исключительно на компьютерный искусственный интеллект нельзя. Итак, задача аналитика, используя медицинские документы и БД как источники атомарных фактов и событий ЛДП, осуществить концептуализацию атомарных фактов и событий. При этом должен возникнуть словарь (тезаурус) предметной области, должна быть, по крайней мере, учтена синонимия и омонимия понятий. В результате должно появиться некоторое множество концептуализированных событий, описывающих знания для определенного домена предметной области.

3. Синтез структуры знаний

Пусть нам дано конечное множество событий D. На каждом событии $S \in D$ на его атомарных фактах задано отношение нестрогого порядка, отражающее контекстные отношения на атомарных фактах из S. Поставим задачу синтеза структуры знаний по заданному множеству событий. Сразу же постулируем нашу структуру в виде дерева атомарных фактов. Более универсальная структура – семантическая сеть, но мы будем рассматривать только дерево. Отчасти это объясняется нашими прагматическими намерениями использовать в качестве языка – носителя структуры – XML. Этот язык уже стал de facto языком описания стандартов обмена медицинскими данными, языком описания структуры медицинских данных.

Начальную вершину искомой структуры будем считать всегда заданной. Задачу синтеза структуры разобьем на две подзадачи.

Задача 1. Для заданной вершины дерева структуры и заданного конечного множества событий $S \in D$ построить множество вершин дочерних по отношению к заданной.

Определим на множестве любых событий отношение эквивалентности $\approx$.

Два события $S^a = \{a^i\}\ u\ S^b = \{b^i\},\quad i \in N,\quad a,b \in A^2$ считаются эквивалентными $S^a \approx S^b$, если $a_1^1 = b_1^1$. Эквивалентны события, у которых совпадают первые абстрактные понятия в первых по порядку атомарных фактах. Отношение эквивалентности порождает разбиение данного множества событий на непересекающиеся классы эквивалентности. Создаем в структуре дерева дочерние вершины по одной для каждого выделенного факторизацией класса эквивалентности. Каждая из созданных дочерних вершин включает в себя атомарный факт, построенный из факторизующего класс абстрактного понятия и пустого конкретного понятия. В результате мы получаем некоторое конечное множество новых вершин и для каждой верши-

ны определено некоторое конечное множество факторизованных событий. Процесс построения структуры следует выполнять итерационно, начиная с начальной вершины и заданного исходного множества событий. После решения для заданной вершины и заданного множества событий задачи 1, из всех разбитых на классы эквивалентности событий исключаем первый по порядку атомарный факт. Исключаем из всех подмножеств событий, связанных с дочерними вершинами, пустые события, появившиеся в результате исключения атомарного факта. Для каждой дочерней вершины и связанного с нею непустого множества событий вновь решаем задачу 1. В силу конечности событий и входящих в них атомарных фактов итерационный процесс сойдется к единственной структуре. О практическом значении выполненного синтеза будет сказано ниже. Отметим сейчас следующий очевидный недостаток приведенного решения. Это решение никоим образом не использует информацию о контекстных отношениях. Можно легко привести пример следующих трех событий (в записи примера опустим конкретные понятия): {(‘a’), (‘b’), (‘c’)}, {(‘b’), (‘c’), (‘a’)}, {(‘c’), (‘a’), (‘b’)}, и будем считать, что во всех трех событиях все атомарные факты в силу заданного отношения порядка равны между собой $a \approx b \approx c$. Очевидно, что мы могли бы путем перестановки атомарных фактов, не нарушая их контекстных отношений, привести все эти три события к виду {(‘a’), (‘b’), (‘c’)}, {(‘a’), (‘b’), (‘c’)}, {(‘a’), (‘b’), (‘c’)}. Построенная по этим событиям структура была бы проще, чем структура, построенная по исходным событиям. Приведенный пример говорит о возможности оптимизации алгоритма синтеза структуры и приведении структуры знаний к некоторому более простому нормализованному виду.

Задача 2. Для заданной вершины дерева структуры и заданного конечного множества событий $S \in D$ построить нормализованное по отношениям порядка множество вершин дочерних по отношению к заданной. Решение задачи 2 сводится к не нарушающим контекстные отношения перестановкам атомарных фактов в заданных событиях, и затем в решении для построения одного уровня структуры задачи 1. Введенное нами на атомарных фактах отношение нестрого порядка $\leq$ определяет на событиях отношение эквивалентности $\approx$

$$a \approx b, \quad \textit{если } a \leq b \quad \textit{и} \quad b \leq a. \tag{5}$$

В силу отношения эквивалентности $\approx$ на атомарных фактах в каждом из заданных событий могут быть построены свои классы эквивалентности. Нас будут интересовать те классы, в которые входят первые по порядку атомарные факты. Пусть Z есть объединение абстрактный понятий из всех классов эквивалентности, в которые входят первые по порядку атомарные факты. Контекстной мощностью абстрактного понятия $z \in Z$ относительно заданного множества событий $S \in D$ назовем число классов эквивалентности, в которые входят атомарные факты с абстрактным понятием $z \in Z$ и одновременно входят любые первые по порядку атомарные факты заданного множества событий. Чем больше контекстная мощность абстрактного понятия, тем чаще оно встречается в контексте первых атомарных фактов и тем выше семантическая значимость этого абстрактного понятия. Формируем последовательность элементов $z \in Z$ по мере убывания их контекстной мощности. Выбираем первый элемент из последовательности и во всех событиях, в которых в классе эквивалентности первого атомарного факта есть атомарный факт с абстрактным понятием, совпадающим с выбранным из последовательности элементом, переставляем атомарный факт с данным абстрактным понятием на первое место. Эта перестановка выполняется в рамках класса эквивалентности и не нарушает контекстных отношений. Исключаем из рассмотрения все события, участвующие в перестановке и исключаем из последовательности первый элемент. Процесс перестановок продолжаем до исчерпания оставшегося в рассмотрении множества событий. Таким образом, семантически более значимые атомарные факты без нарушения контекстных отношений будут поставлены на первое место. Далее для заданной вершины и заданного «нормализованного» множества событий решается задача 1. Решение задачи 1 породит множество дочерних вершин по отношению к заданной и определит для каждой дочерней вершины свое множество далее обрабатываемых событий. После решения задачи 1 исключаем первый по порядку атомарный факт из всех событий в каждом классе эквивалентности. Исключаем из всех подмножеств событий, связанных с дочерними вершинами, пустые события, появившиеся в результате исключения атомарного факта. Для каждой такой дочерней вершины и каждого связанного с нею множества собы-

тий вновь проводим нормализацию и построение новых вершин согласно алгоритму решения задачи 2. Отметим, что приведенный алгоритм построения нормализованной структуры знаний уже не гарантирует единственности решения. Структура будет построена с точностью до перестановок атомарных фактов одной контекстной мощности, не нарушающих контекстные отношения.

Заключение

Остается обсудить значение предлагаемого контекстного анализа событий и синтеза структуры медицинских знаний. Во-первых, предложен простой базис для проведения концептуализации предметной области. В ходе концептуализации должен быть построен словарь или тезаурус понятий предметной области. Концептуализированная, по своей сути процессная, модель предметной области находит свое представление в виде потока событий. Атомарные факты, входящие в события, упорядочиваются по мере их контекстной мощности (семантической значимости). Во-вторых, по множеству концептуализированных событий становится возможным выполнить нормализованный синтез структуры знаний. По алгоритму своего построения структура знаний будет иерархически включать в себя в первую очередь наиболее употребляемые и семантически значимые понятия, с переходом к менее распространенным специализированным понятиям на более удаленных от начальной вершины уровнях иерархии. Например, при построении МИС обычно преобладает подход ориентированный на пациента, а также удовлетворяются требования авторизации всей медицинской информации. Поэтому в структуре знания понятия пациента и автора (медицинского работника) должны подняться на верхние уровни иерархии. Единую медицинскую карту пациента (ЕМК) – можно рассматривать как виртуальный документ, содержащий все события всех ЛДП, связанных с данным пациентом. Возможная структура ЕМК может быть построена с помощью описанного в работе алгоритма синтеза структуры знаний. Аналогичный подход можно распространить и для анализа и построения знания в рамках единого информационного медицинского пространства и для построения структуры мобильной электронной медицинской карты. В настоящее время предложенный подход проходит практическую проверку в рамках МИС Интерин Promis разработанной в исследовательском центре медицинской информатики ИПС РАН. В качестве развития описанного подхода рассматривается возможность введения контекстных отношений на событиях, в частности, темпоральных отношений, с последующим учетом этих отношений при синтезе структуры знания. Рассматривается также возможность синтеза других структур знаний, в частности, семантической сети.

Представленные исследования проводились в рамках гранта № 07-01-00719-а Российского фонда фундаментальных исследований по проекту «Принципы выделения знаний и формирования тезауруса в медицине на основе анализа структуры и содержания электронных медицинских документов».

Список литературы

1. Guliev Y.I., Malykh V.L., Yurchenko S.G. Conceptual models for representing information in healthcare information systems. - Advanced Information and Telemedicine Technologies for Health, AITTH 2005, Minsk. Vol 1, P. 198-201.

2. Гулиев Я. И.-О., Малых В.Л. Архитектура HL-X. // Программные системы: Теория и приложения, том II, стр. 147-168, М.: Физматлит, 2004.