РАБОТА С НЕТОЧНЫМИ ВРЕМЕННЫМИ ДАННЫМИ В СИСТЕМАХ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ

О.С. Бартунов, С.В. Карпов

Современное цифровое общество характеризуется глобальным проникновением различных компьютерных сервисов во все области нашей жизни, будь то развлечение, обучение или научные исследования. Весь цикл научного исследования в той или иной мере связан с компьютерами и разнообразными службами, например, поисковые машины помогают нам разыскивать необходимые данные, архивы статей позволяют обмениваться научными публикациями, а разнообразные сенсоры добывают для нас научные данные. Проникновение цифровых сервисов в научную жизнь не везде одинаково; так, гуманитарные науки только переходят от этапа накопления цифровой информации (перевод в цифровую форму) к этапу ее структурирования и только начали сталкиваться с проблемами хранения в базах данных. Например, исторические науки оперируют неточными датами, которые не поддерживаются современными промышленными СУБД. С другой стороны, и в более продвинутых естественных науках существуют схожие проблемы, связанные с необходимостью работы с неточными данными - например, все экспериментальные данные имеют определенную ошибку и надо уметь это учитывать, надо уметь работать с пропущенными данными и данными с пределами, уметь проводить кросс-идентификацию событий в экспериментах итд. Для специалистов по базам данных все эти проблемы сводятся к необходимости работы со значением, имеющим дополнительный атрибут - интервал значений. Работа означает эффективное хранение таких данных, доступ к ним и поддержку определенных операторов. Мы остановимся на примере неточных временных данных в приложении к историческим наукам, используя наиболее развитую и богатую возможностями свободную СУБД PostgreSQL, которая зачастую используется в научных приложениях.

Основным элементом исторической информации является "событие" - совокупность времени, места, участвующих лиц и собственно содержания события. Знание этих параметров позволяет строить взаимосвязи событий - пространственно-временные, причинные, итд - и делать заключения о деталях исторического процесса. Важными элементами являются также персоналии и библиографии - списки документов, описывающих те или иные события, личности, комментирующие другие источники данных и так далее.

Время события - историческая дата - является зачастую не вполне точно определённой величиной. Действительно, далеко не всегда известно не только точное время того или иного события, но и его день, месяц, год. Более того, зачастую про событие известно лишь то, что оно было до некоторого момента. В целом, можно выделить четыре основных случая такой неопределенности:

- Точная историческая дата - имеет чётко определённую дату и время.
- Неточная дата - известен лишь некоторый интервал, её содержащий. Вот возможные примеры:
  - 1912-02-05 - момент задан с точностью до суток
  - 315 BC - момент с точностью до года
  - XV век - момент с точностью до века
  - во время Реформации - с точностью до длительного интервала
- Дата, заданная правым неравенством - событие, произошедшее до заданного момента, но когда именно - неизвестно
- Дата, заданная левым неравенством - событие после заданного момента

Естественным способом задания подобной информации в СУБД является интервальный тип - в случае PostgreSQL это может быть реализовано либо посредством композитного типа из двух timestamp-ов, либо на основе общего подхода Range Types (http://wiki.postgresql.org/wiki/RangeTypes), который будет доступен, начиная с версии 9.2. Требуемые для формализации работы с данными операции сравнения заданных таким образом моментов времени могут быть заимствованы из интервальной алгебры Аллена и включают в себя следующие отношения:

- **>** и **<** - (**after** и **before**) строго больше или меньше, то есть и начало и конец первого события позже начала второго, или наоборот.
- **>=** и **<=** - (подвиды **overlaps**) условно больше или меньше - то есть начало первого позже начала второго и его конец - тоже после конца второго, и наоборот (фактически асимметричные перекрытия).
- **=** (**during**) равенство для случая, когда один из интервалов полностью лежит внутри другого (неважно, первый или второй) - то есть даты фактически совпадают в пределах неопределенности
- **@>** - (**during**) интервал первой даты полностью покрывает интервал второй

- **<@** - (**during**) интервал первой даты полностью лежит внутри интервала второй
- **&&** - (**overlaps**) просто пересекающиеся интервалами даты (то есть в принципе могли быть одновременно, а могли - и нет).

Таблица основных отношений двух дат, **A** и **B**, заданных моментами начала (**A-**, **B-**) и конца (**A+**, **B+**), приведена ниже:

	A- < B-	B- < A- < B+	A- > B+
A+ < B-	**A < B**	---	---
B- < A+ < B+	**A <= B** **A && B**	**A <= B** **A && B** **A <@ B**	---
A+ > B+	**A = B** **A && B** **A @> B**	**A >= B** **A && B**	**A > B**

Этих отношений достаточно для упорядочивания последовательности событий - построения их "временных последовательностей" (timelines).

Следует заметить, что исторические даты не обязательно задаются непрерывным интервалом - они могут порой сводиться к массиву таких интервалов - временных отрезков. Действительно, неформальное задание момента исторического события вида "осенью XIX века, но скорее всего в сентябре 20-х годов" невозможно свести к одному интервалу - XIX веку - без потери потенциально важной уточняющей информации. Использование массивов интервалов решает эту проблему.

Географическая информация - место, где происходило событие - также не обязательно является точно заданной. Нередки варианты её задания вида "в Москве", "в Германии", "на Ближнем Востоке". СУБД PostgreSQL имеет готовые средства для работы с географической информацией, включающие в себя как полноценные ГИС-решения (PostGIS, http://postgis.refractions.net/), так и более низкоуровневые расширения для работы с координатами на сфере (pgSphere, http://pgsphere.projects.postgresql.org/, и Q3C, http://sourceforge.net/projects/q3c/).

Персоналии задаются списком имён, под которыми историческая личность известна в источниках, а также датами жизни и смерти. Библиографии строятся на основе понятия "документа" - текста, который имеет дату написания и может входить в коллекции-издания. Таблица предикатных связей задаёт различные виды отношений между документами, персоналиями и событиями, которые могут иметь, к примеру, следующий вид:

- историческая личность А - **является автором** - документа Б
- историческая личность А - **упоминается в** - документе Б
- документ А - **ссылается на** - документ Б
- в документе А - **упоминается** - событие Б
- в событии А - **участвовала** - историческая личность Б

Подобная схема позволяет эффективно формализовать, хранить и строить заключения по большинству видов "сырой" информации любого типа, не обязательно описывающей исторические процессы.