Проблемы идентификации участников событий

Д. В. Борисенков, email: xuser@relex.ru

Воронежский государственный университет

Аннотация. *В данной работе рассмотрены типичные проблемы, возникающие при идентификации участников множеств событий в разных предметных областях. Намечены пути обнаружения, решения и предотвращения таких проблем.*

Ключевые слова: *Субъект, событие, участие, идентификация.*

Введение

В различных предметных областях регулярно возникают задачи идентификации субъектов, принимающих участие в множествах однотипных событий. Например, к числу таких предметных областей можно отнести:

- спортивные соревнования, участвующие в них спортсмены и показанные результаты;
- научные конференции, их участники и сделанные доклады;
- метрические книги, содержащие информацию о рождениях, смертях и бракосочетаниях.

Данная статья посвящена рассмотрению типичных проблем, возникающих при идентификации участников таких множеств событий.

Терминология

Субъект – участник множества серийных (однородных) событий. Субъект – родовое понятие, предполагается, что экземпляры типа субъекта – это люди (физические лица).

Субъект характеризуется своими атрибутами. Атрибуты субъекта могут быть разделены на несколько перечисленных ниже категорий.

- Идентификационные атрибуты – каждый такой атрибут представляет собой уникальный идентификатор в некоторой системе идентификации субъектов. Субъект может не иметь идентификационных атрибутов, если система идентификации для данной предметной области не существует (не была создана). Экземпляр субъекта может иметь неопределенное значение идентификационного атрибута, если он не был идентифицирован в соответствующей системе идентификации (идентификация не производилась или не дала результатов).
- Именные атрибуты – фамилия, имя, отчество. Эти атрибуты могут быть указаны частично, сокращенно, в разных алфавитах, с ошибками. В том или ином виде, прямо или косвенно (через

идентификатор в некоторой системе идентификации), заданы для субъекта практически всегда.
- Пространственные атрибуты – место рождения, место проживания или регистрации в момент события. Могут быть указаны страна, область, город, полный адрес.
- Отношение к какой-либо организации – например, федерация для спортсменов, вуз для участников конференций, религиозная конфессия для участников метрических книг. Эта категория атрибутов может частично пересекаться с предыдущей.
- Временные атрибуты – дата или год рождения, возраст в момент события.
- Звания, титулы, разряды – спортивные звания для спортсменов; научные степени, звания, должности для участников конференций; сословия для участников метрических книг.
- Рейтинги (в момент события) – для участников спортивных соревнований.

Атрибуты субъектов могут изменяться во времени (для идентификации участников событий, сильно разнесенных во времени, база данных, содержащая информацию о предметной области, должна иметь хронологическую организацию). В то же время атрибуты субъекта как участника некоторого события являются неизменными.

Событие – мероприятие, обычно имеющее время и место, участником которого являются субъекты. Событиями могут быть: спортивные соревнования; научные конференции, публикации книг и сборников статей; рождения, смерти и бракосочетания, отраженные в метрических книгах. Событие – родовое понятие, которое имеет тип (возможно, с подтипами) и экземпляры.

Для некоторых типов событий может существовать иерархия (спортивное соревнование может состоять из отдельных матчей или партий; отдельные соревнования объединяются в серии; конференция состоит из секций, секции – из докладов, сборник – из статей).

Событие характеризуется своими атрибутами. Атрибуты события могут быть разделены на следующие категории:

- Идентификационные атрибуты – каждый такой атрибут представляет собой уникальный идентификатор в некоторой системе идентификации событий. Событие может не иметь идентификационных атрибутов, если такая система идентификации для данной предметной области не существует (не была создана). Экземпляр события может иметь неопределенное значение идентификационного атрибута, если оно не был идентифицирован в этой системе идентификации

(идентификация не производилась или не дала результатов). В целом идентификация для событий не столь важна, как для субъектов, т.к. информация о событии включает атрибуты участвующих субъектов и атрибуты их участий.

- Именные атрибуты – название события, могут существовать полное и краткое названия. В отличие от субъектов, название события может отсутствовать или не иметь принципиального (в том числе идентификационного) значения.
- Пространственные атрибуты – где произошло событие (частичный или полный адрес, название спортсооружения, вуза, церкви и т.д.). Для составного события такой атрибут может представлять собой совокупность пространственных атрибутов составляющих его событий.
- Временные атрибуты – когда произошло событие (дата и время начала, дата и время окончания). Для составного события такой атрибут может представлять собой – совокупность временных атрибутов составляющих его событий.
- Категория события: для соревнований – турнир, матч, командное соревнование; для конференций – международная, всероссийская, межвузовская; для метрических книг – рождение, смерть, вступление в брак.

Участие – элементарный факт, отражающий задействованность субъекта в событии. Участие – родовое понятие, которое имеет тип и экземпляры.

Участие характеризуется своими атрибутами. Атрибуты участия могут быть разделены на несколько перечисленных ниже категорий.

- Информация о субъекте (инентификационная, именная и прочая).
- Информация о событии (индентификационная, именная и прочая).
- Роль субъекта в событии: для спортивных соревнований – участник, судья, организатор, тренер, капитан команды; для конференций – докладчик, организатор, слушатель; для метрических книг – основной участник (родившийся, умерший, ставіий мужем/женой, мать, отец), свидетель, регистратор.
- Результат участия: для спортивных соревнований – количество сыгранных матчей/партий и набранных очков; для конференций – тема доклада.

Варианты задач идентификации субъектов

Задача создания системы идентификации субъектов. На входе есть первичная информация о группе однородных событий (атрибуты

событий и относящихся к ним субъектов и участий). Предполагается, что среди перечисленных в ней субъектов достаточно велика доля таких, которые принимали участие в нескольких из перечисленных событий. Необходимо на основе этой первичной информации создать систему идентификации участников и сопоставить каждое участие с некоторым идентификатором субъекта.

Задача использования систем идентификации субъектов. Дополнительно к входным данным первой задачи, существует одна или несколько систем идентификации субъектов. Необходимо для каждого участия: либо связать его с некоторым идентификатором субъекта из одной из существующих систем идентификации, либо сделать вывод о том, что подходящего идентификатора для него нет, и создать для таких участий дополнительную систему идентификации.

Задача поиска предполагаемых проблем идентификации в существующей базе данных, содержащей информацию о событиях, субъектах и участиях. На входе есть база данных событий, субъектов и участий, при этом каждое участие отнесено к определенному идентификатору субъекта. Необходимо на основе анализа информации из базы данных сделать выводы о наличии (или отсутствии) в ней проблем идентификации субъектов, перечислив найденные проблемы и дав оценку вероятности для каждой из них.

Варианты проблем идентификации субъектов

Проблема двойников первого типа (мнимый двойник). Один и тот же субъект связан с двумя (или более) разными идентификаторами субъекта. Т.е. участия, в которых в действительности был задействован один и тот же субъект, отнесены в базе данных к разным идентификаторам субъектов.

Проблема двойников второго типа (реальный двойник). Два (или более) разных субъекта связаны с одним и тем же идентификатором субъекта. Т.е. участия, в которых в действительности были задействованы разные субъекты, отнесены в базе данных к одному и тому же идентификатору субъекта.

Неверная идентификация. Существуют два разных субъекта с разными идентификаторами субъекта, у каждого из них в базе данных есть участия, правильно отнесенные к его идентификатору субъекта, но также существует одно (или более) участий одного из этих субъектов, отнесенных в базе данных к идентификатору другого субъекта. Неверная идентификация является комбинацией проблем двойников первого и второго типа.

Источники проблем идентификации субъектов

Основной причиной проблем, перечисленных в предыдущем пункте, является неверная идентификация субъектов по именным атрибутам:

- один зарегистрированный в системе идентификации субъект был принят за другого ввиду сходства значений именных атрибутов;
- зарегистрированный в системе идентификации субъект не был найден ввиду отличий в записи значений именных атрибутов.

Возможны также ошибки при ручном вводе значений идентификационных атрибутов, либо использование значений идентификационных атрибутов из одной системы идентификации для другой.

Исправление проблем идентификации субъектов

Решением проблемы неверной идентификации является замена идентификатора субъекта – атрибута участия (некорректного идентификатора субъекта – корректным). Данное действие является относительно простым, как и обратное ему действие, которое может потребоваться выполнить, если проблема будет впоследствии сочтена ложно диагностированной.

Решением проблемы двойников первого типа (мнимых) является объединение этих двойников, т.е. отнесение всех участий, зарегистрированных на одного из этих идентификаторов субъектов, ко второму. Такое действие является относительно простым и может быть выполнено автоматически, однако требует тщательной проверки своей корректности – подтверждения, что во всех перечисленных участиях в действительности был задействован один и тот же субъект, поскольку обратное действие (разделение двойников) – гораздо сложнее.

Решением проблемы двойников второго типа (действительных) является создание нового идентификатора субъекта для одного из субъектов-двойников и исправление всех ссылок для участий, в действительности относящихся к этому субъекту. Такие действия требуют участия администратора базы данных.

Поиск вероятных проблем идентификации субъектов

Вероятные двойники первого типа (мнимые) часто имеют полное или близкое к полному совпадение по всей совокупности атрибутов (именным, временным, пространственным и прочим), которые не являются для них неопределенными.

Вероятные двойники второго типа (действительные) также обычно имеют значительное совпадение по большей части атрибутов, не

являющихся неопределенными, однако могут иметь и существенные отличия по некоторым атрибутам. То же самое относится и к случаям неверной идентификации.

Для поиска всех типов проблем полезно также установление неявных связей между сущностями базы данных (например, участие либо неучастие некоторых множеств субъектов в одних и тех же множествах событий).

Заключение

В настоящее время для решения нескольких актуальных задач идентификации субъектов, относящихся к перечисленным в статье категориям, планируется использовать аппарат нейронных сетей, реализованный в виде свободно распространяемых библиотек для языка Python. В качестве подзадач рассматриваются приведение данных для конкретных предметных областей к виду, пригодному для использования в нейронных сетях, выбор архитектуры нейронных сетей и настройка их параметров. Также планируется продолжить анализ различных проблем идентификации субъектов, часть которых рассмотрена в [1-5].

Литература

1. Двойники. Объединение. Модерация. [Электронный ресурс]: база данных. Режим доступа: https://wiki.is-mis.ru/pages/viewpage.action?pageId=54429982

2. Добрякова Г.Э. Проблемы идентификации в информационных системах./ Г.Э.Добрякова. // Правовая информатика. – 2014. – №3. – С. 28-32.

3. Kwon Y. Identifying and removing duplicate records from systematic review searches / Y. Kwon, M. Lemieux, J. McTavish, N. Wathen. // Journal of the Medical Library Association. – Oct 2015. – Vol. 103. – Issue 4. – P.184-188.

4. Strategies around handling duplicate contacts and merging. Documentation web site for technology best practices. [Электронный ресурс]: Режим доступа: https://network.progressivetech.org/node/1158

5. Хабр. Боремся с дубликатами. [Электронный ресурс]. Режим доступа: https://habr.com/ru/post/179789/