ПРОБЛЕМЫ ИДЕНТИФИКАЦИИ УЧАСТНИКОВ СОБЫТИЙ

Борисенков Д.В.

Кафедра МО ЭВМ факультета ПММ ВГУ

ПРЕДМЕТНЫЕ ОБЛАСТИ, В КОТОРЫХ ВОЗНИКАЮТ ЗАДАЧИ ИДЕНТИФИКАЦИИ УЧАСТНИКОВ

- Спортивные соревнования, их участники, тренеры участников, судьи, организаторы соревнований, результаты участников
- Научные конференции, их участники, организаторы и слушатели, доклады и публикации участников
- События, отраженные в метрических книгах (рождения, смерти, бракосочетания), их участники, свидетели и регистраторы
- …

ИДЕЯ ИССЛЕДОВАНИЯ

- Типичный вопрос: в двух разных событиях принимали участие субъекты со сходными атрибутами – это один и тот же реальный субъект или разные?
- Не всегда достаточно информации, для того, чтобы ответить на такой вопрос
- Если ответ на такой вопрос был дан, то нет гарантии, что он не оказался ошибочным
- Для каждой из указанных предметных областей проблемы идентификации участников событий решаются по-разному
- Не удалось обнаружить элементов общего подхода к решению таких задач

ТЕРМИНОЛОГИЯ

- Субъект (сильная сущность)
- Событие (сильная сущность)
- Участие (слабая сущность – каждое участие должно быть связано с одним субъектом и одним событием)

ЭЛЕМЕНТ ЗАДАЧИ – СУБЪЕКТ

- Участник множества серийных (однородных) событий
- Субъект характеризуется своими атрибутами
- Основная задача – идентификация субъектов (чтобы одному субъекту в реальной предметной области соответствовал один субъект в базе данных)
- Недостаточность информации приводит к проблемам идентификации субъектов, в некоторых случаях к невозможности идентификации

РЕАЛЬНЫЕ ВОПРОСЫ ИДЕНТИФИКАЦИИ СУБЪЕКТОВ – УЧАСТНИКОВ СОБЫТИЙ

- В метрических книгах одной и той же деревни неоднократно упоминаются **Савелий Мартынович** и **Савва Мартынович**. Может ли это быть один и тот же человек или это разные люди (например, два брата)?
- В турнирах в некотором шахматном клубе неоднократно участвовал **Синяговский Игорь** с известной датой рождения и один раз – **Снеговский Игорь** с той же датой рождения. Это один и тот же участник и во втором случае ошибка в фамилии?
- В отчете о соревнованиях в Борисоглебске указано, что там принимал участие **Олег Титов из Воронежа**. Но в базе данных нет сведений, что он ранее когда-либо принимал участие в соревнованиях в Борисоглебске, зато несколько раз там выступал его полный тезка – **Олег Титов из Борисоглебска**. Нет ли ошибки в отчете?

АТРИБУТЫ СУБЪЕКТА

- Идентификационные (номера в некоторых системах идентификации)
- Именные (ФИО)
- Пространственные – место рождения, место проживания/регистрации
- Отношение к какой-либо организации
- Временные – дата или год рождения, возраст в момент события
- Звания, титулы, разряды и т.п.
- Числовые характеристики – рейтинги спортсменов, индекс Хирша и т.п. для ученых

ЭЛЕМЕНТ ЗАДАЧИ – СОБЫТИЕ

- Мероприятие, обычно имеющее время и место, участниками которого являются субъекты
- Для некоторых типов событий может существовать иерархия

АТРИБУТЫ СОБЫТИЯ

- Идентификационные (номера в некоторых системах идентификации)
- Название события (может отсутствовать или не иметь значения)
- Пространственные – где произошло событие. Для составного события – совокупность пространственных атрибутов составляющих событий
- Временные – когда произошло событие (начало, окончание). Для составного события – совокупность временных атрибутов составляющих событий
- Категория события – своя для каждой предметной области

ЭЛЕМЕНТ ЗАДАЧИ – УЧАСТИЕ

- Элементарный факт, отражающий задействованность субъекта в событии
- Характеризуется связями с субъектом и с событием, а также атрибутами участия

АТРИБУТЫ УЧАСТИЯ

- Информация о субъекте
- Информация о событии
- Роль субъекта в событии (зависит от предметной области)
- Результат участия (для некоторых предметных областей)

ЗАДАЧА СОЗДАНИЯ СИСТЕМЫ ИДЕНТИФИКАЦИИ СУБЪЕКТОВ

- На входе – первичная информация о группе однородных событий (атрибуты событий и относящихся к ним субъектов и участий)
- Предполагается, что среди упомянутых 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 предыдущей задачи, существует одна или несколько систем идентификации субъектов. Необходимо для каждого участия выбрать один из вариантов:

1) связать его с некоторым идентификатором субъекта из одной из существующих систем идентификации;
2) предложить несколько на выбор несколько возможных идентификаторов субъекта (указав также для каждого из них вероятность того, что он подходит, и, соответственно, вероятность того, что не подходит ни один из них);
3) сделать вывод о том, что подходящего идентификатора субъекта для участия нет

- Создать для таких участий, для которых подходящих идентификаторов не нашлось или они были отвергнуты, дополнительную систему идентификации

ЗАДАЧА ПОИСКА ПРЕДПОЛАГАЕМЫХ ПРОБЛЕМ ИДЕНТИФИКАЦИИ В СУЩЕСТВУЮЩЕЙ БАЗЕ ДАННЫХ

- На входе – 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

ДВОЙНИКИ ПЕРВОГО РОДА

- Одному и тому же реальному субъекту соответствуют два (или более) различных идентификаторов субъектов в базе данных
- Таким образом, участия, в которых в действительности был задействован один и тот же реальный субъект, отнесены в базе данных к разным идентификаторам субъектов

УСТРАНЕНИЕ ДВОЙНИКОВ ПЕРВОГО РОДА (ОБЪЕДИНЕНИЕ)

- Решением проблемы двойников первого рода является объединение этих двойников, т.е. отнесение всех участий, зарегистрированных на одного из этих идентификаторов субъектов, ко второму (или наоборот)
- 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

ДВОЙНИКИ ВТОРОГО РОДА

- Два (или более) разных реальных субъекта связаны с одним и тем же идентификатором субъекта в базе данных
- Таким образом, участия, в которых в действительности были задействованы разные реальные субъекты, отнесены в базе данных к одному и тому же идентификатору субъекта

УСТРАНЕНИЕ ДВОЙНИКОВ ВТОРОГО РОДА (РАЗДЕЛЕНИЕ)

- Решением проблемы двойников второго рода 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

НЕВЕРНАЯ ИДЕНТИФИКАЦИЯ

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

УСТРАНЕНИЕ ПОСЛЕДСТВИЙ НЕВЕРНОЙ ИДЕНТИФИКАЦИИ

- 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

ИСТОЧНИКИ ПРОБЛЕМ ИДЕНТИФИКАЦИИ СУБЪЕКТОВ

- Один зарегистрированный в системе идентификации субъект был принят за другого ввиду сходства значений именных атрибутов
- Незарегистрированный в системе идентификации субъект был принят за зарегистрированного ввиду сходства значений именных атрибутов
- Зарегистрированный в системе идентификации субъект не был найден ввиду отличий в записи значений именных атрибутов
- Были допущены ошибки при ручном вводе значений идентификационных атрибутов
- Значение идентификационного атрибута из одной системы идентификации были использовано для другой

ПОИСК ВЕРОЯТНЫХ ПРОБЛЕМ ИДЕНТИФИКАЦИИ СУБЪЕКТОВ

- Вероятные двойники первого рода часто имеют полное или близкое к полному совпадение по всей совокупности атрибутов, которые не являются для них неопределенными
- Вероятные двойники второго рода 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 установление неявных связей между сущностями базы данных (например, участие либо неучастие некоторых множеств субъектов в одних и тех же множествах событий)

ИНСТРУМЕНТАРИЙ ДЛЯ РЕШЕНИЯ ПОСТАВЛЕННОЙ ЗАДАЧИ

- Аппарат нейронных сетей (библиотеки для языка Python)
- Предварительное приведение данных для конкретных предметных областей к виду, пригодному для использования в нейронных сетях

СПАСИБО ЗА ВНИМАНИЕ!