ЮУРГУ
ВШ ЭКН
Кафедра «Защита информации»

Модель нарушителя в системах обезличенных персональных данных

Мищенко Е.Ю., аспирант
научный руководитель Соколов А.Н., к.т.н.

Метод изменения состава или семантики

Реальные данные

Фамилия	Имя	Отчество	Улица	Дом	Квартира	Прочее

Обезличенные данные (варианты)

Квартира	Имя	Фамилия	Улица	Дом	Отчество	Прочее
Фамилия	Ями	Отчество	Улица	Дмо	Триаквра	Прочее
Ф9мэлэя	Эмя	(тчеств(	Улэц9	Д(м	Кв9ртэр9	Прочее

Метод перемешивания

Реальные данные

1	Фамилия	Имя	Отчество	Улица	Дом	Квартира	Прочее
2	Фамилия	Имя	Отчество	Улица	Дом	Квартира	Прочее
.	…	…	…	…	…	…	…
x	Фамилия	Имя	Отчество	Улица	Дом	Квартира	Прочее

Обезличенные данные

1	Фамилия	Имя	Отчество	Улица	Дом	*7артира	Прочее
2	Фамилия	Имя	Отчество	Улица	Дом	Ё!артира	Прочее
.	…	…	…	…	…	…	…
x	Фамилия	Имя	Отчество	Улица	Дом	Х.артира	Прочее

Модель нарушителя

Знакомые лица

Фамилия	Имя	Отчество	Улица	Дом	Квартира	1Знакомый
Фамилия	Имя	Отчество	Улица	Дом	Квартира	2Знакомый

Алгоритм кодирования

Квартира	Дом	Фамилия	Улица	Имя	Отчество	2Знакомый
Квартира	Дом	Фамилия	Улица	Имя	Отчество	Прочее
Квартира	Дом	Фамилия	Улица	Имя	Отчество	2Знакомый

Какой из?

Что это?

Что это?

Обезличенные данные

Модель нарушителя

Искомое лицо

Фамилия	Имя	Отчество	Улица	Дом	Квартира	1Искомый
Фамилия	Имя	Отчество	Улица	Дом	Квартира	2Искомый

раскодирование

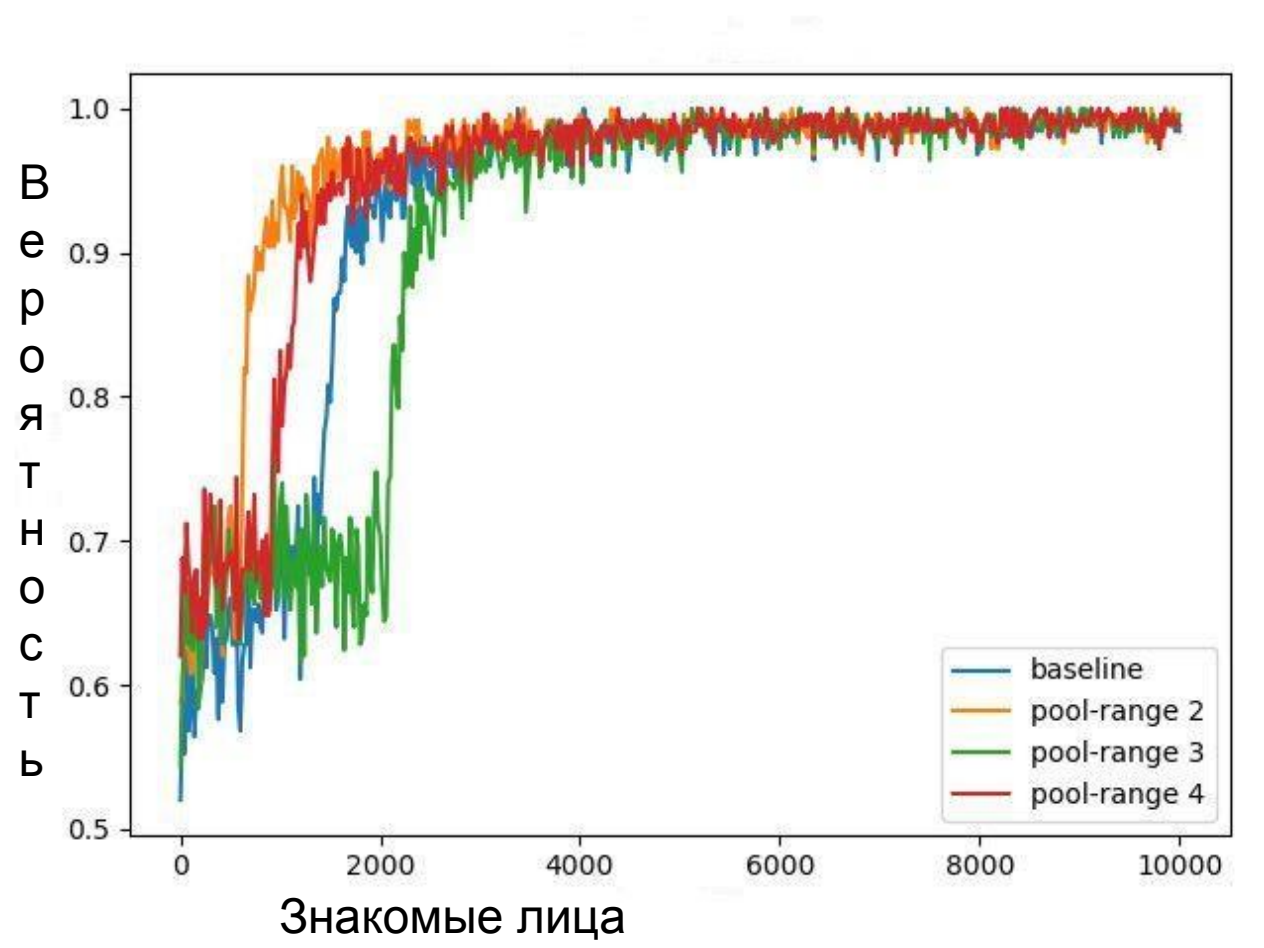
H1	H2	H3	H4	H5	H6	2-й слой
H1	H2	H3	H4	H5	H6	1-й слой
Квартира	Дом	Фамилия	Улица	Имя	Отчество	2Знакомый
Квартира	Дом	Фамилия	Улица	Имя	Отчество	1Знакомый
Квартира	Дом	Фамилия	Улица	Имя	Отчество	Прочее

Обезличенные данные

Метод изменения состава или семантики

- **Объем обработки ПД – 310 тыс. физ.лиц**
- **В базе хранятся следующие ПД:**
  Идентификаторы:
  - ФИО (обрезаны до одной буквы)
  - адрес (перетасованы поля дом и квартира)
  Прочие данные:
  - телефон (уникальный)

Результаты эксперимента



Результаты эксперимента

1) Вероятность раскодирования – 0,9 (500 шагов)
2) Вероятность ошибки по прочим данным – 0,5 (МЕЮ)
3) Перетасовка полей - ненадежна

Выводы

1) Чем больше объем базы – тем больше знакомых лиц
2) Вероятность идентификации знакомых по прочим данным может быть мала. Нужна предварительная обработка (предобучение нейросети)
3) Вероятность раскодирования может быть недостаточна. Нужна оптимизация (фильтры):

- частотный анализ идентифицирующих полей (ФИО, адрес)
- лингвистический анализ сочетаемости букв в текстовых атрибутах (ФИО, адрес)

Спасибо за внимание

Вопросы?