УДК 004.624

Сопоставление информационных объектов из разных баз данных

Солдатенков Роман Владимирович
филиал «НИУ «МЭИ» в г.Смоленске
студент

Аннотация
Предложен алгоритм сопоставления информационных объектов из разных баз данных при совпадении их семантики, но различии синтаксиса выбранных полей. Представлено программное обеспечение, реализующее предложенных алгоритм.
Ключевые слова: сопоставление информационных объектов, базы данных

Comparison of information objects from different databases

Soldatenkov Roman Vladimirovich
National Research University «Moscow Power Engineering Institute» (Branch) in Smolensk
Student

Abstract
An algorithm is proposed for comparing information objects from different databases when their semantics coincide, but the syntax of the selected fields differs. Presented software that implements the proposed algorithm.
Keywords: mapping information objects, databases

Переход экономики на цифровые методы поддержки бизнес-процессов приводит к быстрому росту объема хранимой информации, как в органах государственной власти, так и в коммерческих и общественных организациях. Наблюдавшееся до последнего времени отсутствие строгой стандартизации формы представления сведений об информационных объектах приводило к тому, что одни и те же данные о нем имели различный синтаксис, хотя и содержали одинаковую смысловую (семантическую) нагрузку.

При таком не согласованном формате данных возникают трудности при сопоставлении информации из различных источников, в частности из баз данных (БД) разных организаций хранящих сведения об одних и тех же объектах. Потребность в таком сопоставлении возникает все чаще и чаще, например, при интеграции сервисов различных структурных подразделений одной компании; при необходимости реализации идеологии «одного окна» при предоставлении государственных услуг в многофункциональных центрах; при анализе сведений о субъекте, хранящихся в базах, сформированных в сильно разнесенные по временной оси периоды и,

соответственно, имеющие разные формы и стандарты представления данных. Указанные обстоятельства могут свидетельствовать об актуальности разработки методов, позволяющих осуществлять подобные сравнения БД.

Реализация сопоставления записей в обозначенных условиях на практике сталкивается как с несоответствием количества полей в БД для обозначения одного и того же атрибута, так и с ошибочным или сокращенным его написанием. Для человека нет проблемы для идентификации записей с такими атрибутами из разных баз данных, но для вычислительной машины эта задача является не тривиальной.

В большинстве СУБД есть специальные функции которые могут осуществить сопоставление разных баз данных по значениям каких либо полей (например, операция inner join в MS Access), но если при этом не совпадает хотя бы один символ, такие операции приводят к результату «ложь», то есть содержимое поля считается не совпадающим, хотя записи, по смыслу, содержат одну и ту же информацию. Таким образом, допускается ошибка первого рода – принимаем отрицательное решение для ситуации, когда должно быть принято положительное.

Автором предложен эмпирический метод сопоставления записей из разных БД, относящихся к одному объекту. Работа выполнялась по заказу страховой компании, которой было необходимо объединить данные об одних и тех же клиентах, но хранящихся в базах, как самой компании, так и в базах смежных учреждений (в частности, налоговой службы). Такой симбиоз позволил бы страховщику учитывать ряд дополнительных признаков клиента (из других БД), которые не прописываются в страховом договоре, но которые могли бы более качественно и взвешенно принимать решение о тарифах или, даже об отказе заключения договора. Принятие этого решения базируется на выполняемом после сопоставления выявлении факторов (полей, описывающих клиента), которые в наибольшей степени характерны для мошенников. В своей постановке эта задача может быть отнесена к обратным, так как по наблюдаемому выходу системы (размеру страховой выплаты) требуется установить причины, вызвавшие его. Методы решения подобных задач требуют наличия аналитических описаний предметной области или других ее моделей, создание которых является весьма трудоемким процессом [1]. Поэтому была сделана попытка найти более простой в реализации вариант решения.

Отметим, что актуальность разработки подтверждается тенденциями последнего времени на рынке страхования – увеличением числа имитаций дорожно-транспортных происшествий и применением мошеннических схем получения выплат по ОСАГО. Так, «АльфаСтрахование» в 2017 году подало около 1 тыс. заявлений на подобные случаи, что почти на 30% больше по сравнению с прошлым годом. В результате правоохранительными органами были возбуждены 104 уголовных дела против 48 годом ранее [2].

В основе предлагаемого метода сопоставления записей БД лежит подсчет количества повторений одинаковых символов в анализируемых полях записей и формирование для анализируемого клиента сигнатуры или

вектора-строки S=[s_1, s_2 , … s_{42}], каждый элемента которого равен числу повторений соответствующего символа алфавита (s_1 соответствует числу повторений буквы А, s_{32} – буквы Я, s_{33} – цифры 0, s_{42} – цифры 9). Количества точек, запятых и других символов не производится, так как их число обычно относительно не велико. Для предотвращения разночтения с прописными и строчными буквами все символы переводились в строчные. Кроме этого буква Ё приравнивалась к букве Е. При реализации алгоритма было учтено, что в большинстве БД клиенты в обязательном порядке заносят свои имя, фамилию и отчество, а также адрес, поэтому именно данные поля использовались для контроля совпадения.

Блок-схема алгоритма сопоставления полей для одного клиента имеет описание, представленное ниже.

Начало.

1. F1=ФИО клиента из БД1.
2. F2=ФИО клиента из БД2.
2. F1=F2? Если «Нет», pr = 0 и переход к п. 5.
3. Расчет сигнатур S1 и S2 для клиента по данным из полей «Адрес»из БД1 и БД2 соответственно.
4. S1=S2? Если да, то pr=1, в противном случае pr=0.
5. Конец.

Результатом работы алгоритма является признак pr, который принимает значение «истина» при совпадении записей и «ложь» - в противном случае. Эта информация применяется в дальнейшем для формирования обобщенной записи, содержащей поля из разных БД. Кроме указанных укрупненных шагов алгоритм еще предусматривает возможность сопоставления имен клиентов при указании их как в одном, так и в трех разных полях (отдельно для имени, фамилии и отчества).

Остановимся более подробно на реализации пункта 4 представленного алгоритма. Полного равенство сигнатур добиваться не следует, так как, например, одну и ту же улицу можно записывать как сокращенно, так и в полном виде (ул. Маршала Соколовского или ул. М. Соколовского), но тем не менее это будет одна и та же улица. Поэтому была предусмотрена возможность ввода количества элементов сигнатур S1 и S2, при совпадении которых считалось, что поля адресов совпадают.

Программной средой для реализации программы сопоставления БД выступал MS Excel. Это выбор был обусловлен тем, что большинство имеющихся в страховой компании БД были выполнены именно в форме электронных таблиц. Также, в техническом задании было указано, что если БД были бы созданы в других СУБД, то отдел по информатизации предоставил бы их копии, перенесенные в формат Excel. Наличие в составе MS Excel встроенного языка программирования VBA позволило разработать развитую программу, обладающую удобным и интуитивно понятным оконным интерфейсом, выполняющую не только сопоставление, но и дальнейший корреляционный анализ данных. Такой анализ проводился для полей уже объединенной БД, что позволяло выявить статистически

значимую зависимость между содержимым выбранных полей и величиной выплачиваемой суммы, а следовательно, сделать суждение о возможности применения клиентом мошеннических схем. Главная форма программы показана на рисунке 1.

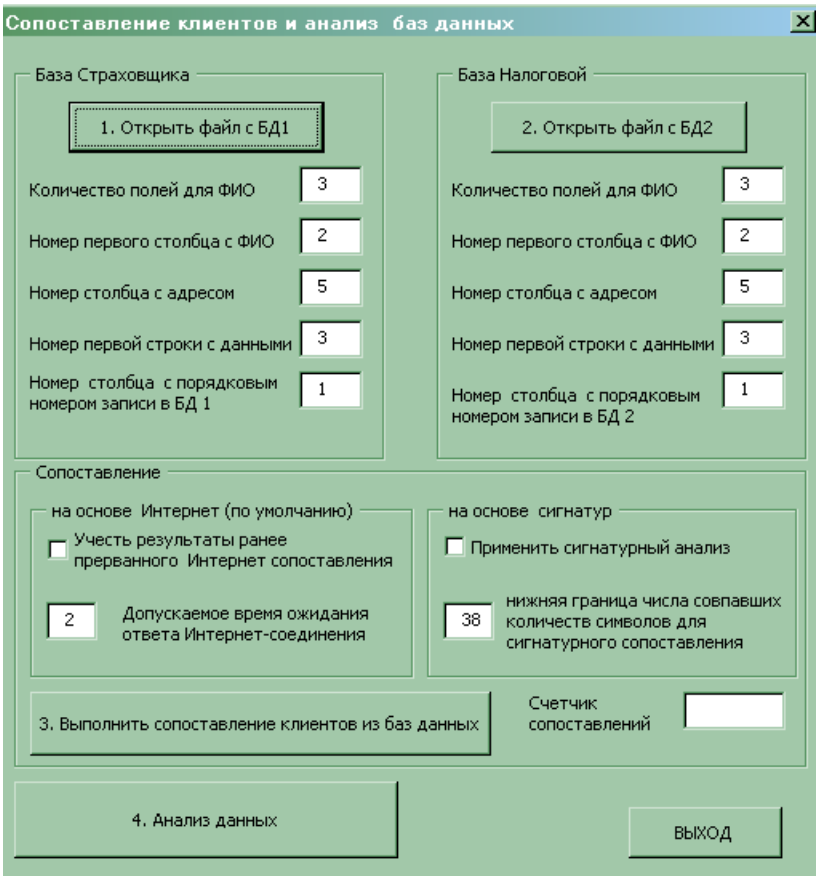
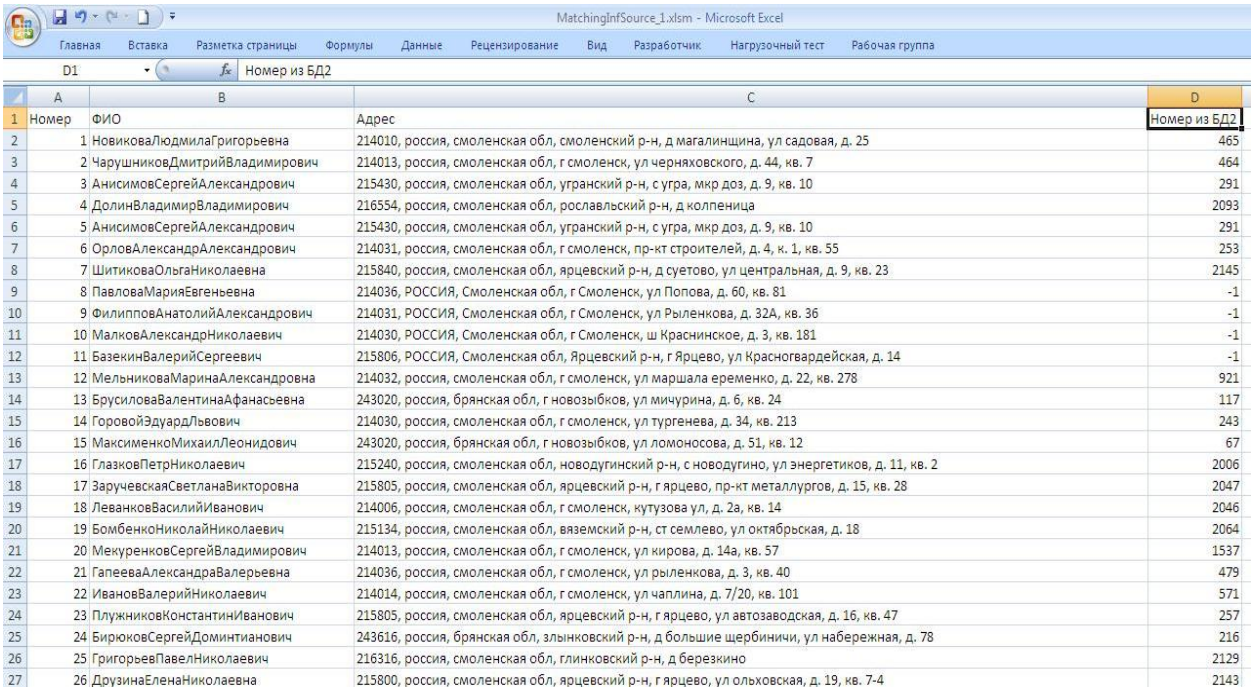


Рисунок 1 – Главная форма программы

Расчетное обоснование предложенного метода сопоставления полей записей БД с точки зрения статистики не проводился, но, очевидно, что доля ошибочных решений может присутствовать при совпадении имен и небольшом расхождении в написании адресов. Однако, если ошибки и возможны, то их общая доля, как предполагалась, будет мала и не повлияет на общий результат дальнейшего статистического анализа. Это предположение было подтверждено после обработки существующей БД клиентов, содержащей более 40 000 записей – ни одного ошибочного сопоставления выявлено не было, так как на практике очень редко можно встретить полных однофамильцев, проживающих рядом. Выходные данные программы формируются в виде таблицы Excel. Фрагмент выходной формы программы показан на рисунке 2.

На рисунке 2 в крайнем правом столбце указан номер строки, в которой она представлена в другой базе данных. Значение «-1» отражает тот факт, что строка не найдена, то есть, нет соответствия. На основании этой информации алгоритм объединяет выбранные поля из первой и второй баз данных. Это

позволяет проводить более глубокий статистический анализ клиента на основании объединенной информации.



MatchingInfSource_1.xlsm - Microsoft Excel

Главная | Вставка | Разметка страницы | Формулы | Данные | Рецензирование | Вид | Разработчик | Нагрузочный тест | Рабочая группа

D1 | Номер из БД2

	A	B	C	D
1	Номер	ФИО	Адрес	Номер из БД2
2	1	НовиковаЛюдмилаГригорьевна	214010, россия, смоленская обл, смоленский р-н, д магалинщина, ул садовая, д. 25	465
3	2	ЧарушниковДмитрийВладимирович	214013, россия, смоленская обл, г смоленск, ул черняховского, д. 44, кв. 7	464
4	3	АнисимовСергейАлександрович	215430, россия, смоленская обл, угранский р-н, с угра, мкр доз, д. 9, кв. 10	291
5	4	ДолинВладимирВладимирович	216554, россия, смоленская обл, рославльский р-н, д колпеница	2093
6	5	АнисимовСергейАлександрович	215430, россия, смоленская обл, угранский р-н, с угра, мкр доз, д. 9, кв. 10	291
7	6	ОрловАлександрАлександрович	214031, россия, смоленская обл, г смоленск, пр-кт строителей, д. 4, к. 1, кв. 55	253
8	7	ШитиковаОльгаНиколаевна	215840, россия, смоленская обл, ярцевский р-н, д суетово, ул центральная, д. 9, кв. 23	2145
9	8	ПавловаМарияЕвгеньевна	214036, РОССИЯ, Смоленская обл, г Смоленск, ул Попова, д. 60, кв. 81	-1
10	9	ФилипповАнатолийАлександрович	214031, РОССИЯ, Смоленская обл, г Смоленск, ул Рыленкова, д. 32А, кв. 36	-1
11	10	МалковАлександрНиколаевич	214030, РОССИЯ, Смоленская обл, г Смоленск, ш Краснинское, д. 3, кв. 181	-1
12	11	БазекинВалерийСергеевич	215806, РОССИЯ, Смоленская обл, Ярцевский р-н, г Ярцево, ул Красногвардейская, д. 14	-1
13	12	МельниковаМаринаАлександровна	214032, россия, смоленская обл, г смоленск, ул маршала еременко, д. 22, кв. 278	921
14	13	БрусиловаВалентинаАфанасьевна	243020, россия, брянская обл, г новозыбков, ул мичурина, д. 6, кв. 24	117
15	14	ГоровойЭдуардЛьвович	214030, россия, смоленская обл, г смоленск, ул тургенева, д. 34, кв. 213	243
16	15	МаксименкоМихаилЛеонидович	243020, россия, брянская обл, г новозыбков, ул ломоносова, д. 51, кв. 12	67
17	16	ГлазковПетрНиколаевич	215240, россия, смоленская обл, новодугинский р-н, с новодугино, ул энергетиков, д. 11, кв. 2	2006
18	17	ЗаручевскаяСветланаВикторовна	215805, россия, смоленская обл, ярцевский р-н, г ярцево, пр-кт металлургов, д. 15, кв. 28	2047
19	18	ЛеванковВасилийИванович	214006, россия, смоленская обл, г смоленск, кутузова ул, д. 2а, кв. 14	2046
20	19	БомбенкоНиколайНиколаевич	215134, россия, смоленская обл, вяземский р-н, ст семлево, ул октябрьская, д. 18	2064
21	20	МекуренковСергейВладимирович	214013, россия, смоленская обл, г смоленск, ул кирова, д. 14а, кв. 57	1537
22	21	ГапееваАлександраВалерьевна	214036, россия, смоленская обл, г смоленск, ул рыленкова, д. 3, кв. 40	479
23	22	ИвановВалерийНиколаевич	214014, россия, смоленская обл, г смоленск, ул чаплина, д. 7/20, кв. 101	571
24	23	ПлужниковКонстантинИванович	215805, россия, смоленская обл, ярцевский р-н, г ярцево, ул автозаводская, д. 16, кв. 47	257
25	24	БирюковСергейДоминтианович	243616, россия, брянская обл, злынковский р-н, д большие щербиничи, ул набережная, д. 78	216
26	25	ГригорьевПавелНиколаевич	216316, россия, смоленская обл, глинковский р-н, д березкино	2129
27	26	ДрузинаЕленаНиколаевна	215800, россия, смоленская обл, ярцевский р-н, г ярцево, ул ольховская, д. 19, кв. 7-4	2143

Рисунок 2 – Выходная форма программы

Предложенный алгоритм, а также его программная реализация могут найти применение во многих задачах обработки информации содержащейся в разных БД, когда требуется проводить сравнение информационных объектов по характеристикам, которые могут записываться с отклонениями в синтаксисе, но иметь одинаковое семантическое значение.

Библиографический список

1. Пучков А.Ю., Павлов Д.А. Алгоритмы поиска решения обратных задач при непрерывном и дискретном времени // Научное обозрение. 2013. № 1. С. 174-176.
2. Мошенничество пошло в рост // РБК. Газета. № 023 (2747)(0802). 07.02.2018. [электронный ресурс]. URL: https://www.rbc.ru/newspaper/2018/02/08/5a7b19cc9a794795fc9df320 (дата обращения 01.03.2018).