Ю. П. Кулябичев, Р. В. Коновалов

ПОВЫШЕНИЕ БЕЗОПАСНОСТИ АВИАПЕРЕВОЗОК ПУТЕМ СОЗДАНИЯ СИСТЕМЫ КОНТРОЛЯ СПИСКОВ ПАССАЖИРОВ

Преамбула

После событий 11 сентября 2001 г., произошедших в США, были существенно повышены требования к безопасности в гражданской авиации. В последующие несколько лет на международных рейсах США была создана многоуровневая система контроля пассажиров, препятствующая попаданию на борт воздушного судна нежелательных лиц. Позже элементы этой системы были внедрены на международных рейсах Европейского союза. В течение последующего десятилетия высокие требования к безопасности, внедренные западными странами, и системы контроля распространились практически по всему миру.

Одним из элементов новой системы безопасности в пассажирской авиации является контроль пассажиров с целью сравнения с различными списками. Условно в дальнейшем будем называть такой список *списком нежелательных лиц*, хотя недопущение на рейс может не являться целью выявления тех или иных субъектов в общем пассажиропотоке.

В статье рассматривается задача сверки пассажиров авиационного рейса со списком нежелательных лиц путем сравнения имен. Сравнение паспортных данных не несет технических сложностей, но не всегда возможно в силу специфики процедуры бронирования в коммерческой авиации. Эти особенности будут рассмотрены ниже. Задача сравнения имен, в отличие от проверки паспортов, требует не выявления точного совпадения, а использования алгоритмов приближенного поиска в силу национальных особенностей написания и возможных искажений имен.

1. Проблемная и предметная области

Перед посадкой в самолет пассажир проходит через ряд процедур, из которых со стороны авиакомпании-перевозчика можно выделить три: бронирование, покупку билета и регистрацию на рейс. Пассажир может приехать в аэропорт непосредственно перед вылетом рейса без каких-либо предварительных процедур и, если на данный вылет есть свободные места, купить билет и зарегистрироваться на рейс. Однако в большинстве случаев три указанные процедуры разнесены во времени.

Сначала пассажир, найдя приемлемый для него маршрут, время перевозки, авиакомпанию и тариф, осуществляет бронирование места на рейс. Далее ему необходимо оформить билет и произвести оплату. При этих процедурах данные поступают в так называемую систему бронирования. Такие системы в рамках национальных реализаций есть у небольшого количества

стран. В основном авиакомпании используют международные системы, которые называются глобальными дистрибутивными системами (ГДС, или global distribution systems, GDS) [1].

До 2001 г. процедура бронирования предназначалась для отслеживания заполнения рейса. Она не являлась юридическим основанием для осуществления перевозки пассажира авиакомпанией (в отличие от оформленного и оплаченного билета). Поэтому в рамках процедуры бронирования в ГДС заносилось только имя пассажира, что являлось достаточным для его идентификации, поиска ранее созданной брони и выписки билета. В рамках процедуры оформления билета паспортные данные пассажира также в ГДС не сохранялись. По состоянию на настоящее время, согласно требованию Министерства транспорта РФ, уже на этапе процедуры бронирования в ГДС заносятся данные документа, удостоверяющего личность [2].

Тем не менее паспортные данные пассажира в ГДС могут отсутствовать. Для максимизации продаж бронирование и оформление билетов авиакомпаниями осуществляется не только через собственные кассы и офисы, но также через агентскую сеть. Например, у крупнейшего российского авиаперевозчика ОАО «Аэрофлот — российские авиалинии» по состоянию на 2006 г. доля собственных продаж составляла 15 % [3]. Таким образом, на авиакомпанию, осуществляющую международные перевозки, бронированием может заниматься какой-либо из агентов, зарегистрированных в различных точках мира и находящихся в поле регулирования местного законодательства. В свою очередь агент может осуществить бронирование по различным каналам, например по телефону, без визуального контакта с пассажиром. Поэтому после прохождения процедуры бронирования (особенно в случае, если оформление билета еще не было произведено) в ГДС может присутствовать только имя пассажира, а паспортные данные отсутствовать.

В аэропорту перед вылетом рейса осуществляется процедура регистрации пассажиров. Данные при этом сохраняются в системе управления отправками аэропорта (Departure Control System, DCS) [1]. Из этой системы информация о пассажирах поступает в ГДС и системы других аэропортов.

Люди могут по каким-либо причинам утрачивать, обменивать и оформлять новые удостоверяющие личность документы международного образца. При этом может возникать временной лаг между обновлениями информации по паспортам в списке нежелательных лиц и появлением пассажира на любом из международных рейсов, осуществляемых различными авиакомпаниями. Наконец, в зависимости от источника информации в списке нежелательных лиц могут отсутствовать паспортные данные.

Таким образом, перед различными организациями, осуществляющими или контролирующими авиаперевозки пассажиров, стоит задача сверки списков нежелательных лиц со списками из ГДС (заранее, до вылета рейса) или из DCS (во время процедуры регистрации) путем сравнения имен. Во втором случае предъявляются требования к скорости обработки списков — все процедуры контроля должны быть осуществлены до посадки пассажиров в самолет.

2. Постановка задачи поиска по сходству

Можно выделить следующие основные причины необходимости учета искажений при проверке пассажирских списков:

• ошибки ввода имен при осуществлении бронирования мест на рейс;

• в системах ГДС и DCS используются латинские буквы, но язык ввода может быть или английский, или французский, так как в паспортах международного образца равнозначно применяются оба этих языка;

• особенности национальных языков, в ряд звуков которых не соответствует ни английскому, ни французскому;

• отсутствие устоявшихся правил трансляции национальных фамилий того или иного региона мира в английскую или французскую транскрипцию.

Задача сверки имен с учетом возможных искажений относится к классу задач *поиска по сходству* или *нечеткого поиска*. Для ее решения существуют различные стандартные алгоритмы (см., например, [4]). Также могут применяться лингвистические подходы.

3. Количество срабатываний алгоритма

Применение при сравнении имен алгоритмов приближенного поиска предполагает выдачу на выходе результатов предположительного совпадения. Эти результаты требуют анализа и соответствующего принятия решений специалистами, осуществляющими процедуры контроля пассажиропотока. Будем называть каждый случай совпадения при сравнении имен *срабатыванием алгоритма*. Поиск по сходству предопределяется количеством срабатываний алгоритма.

Допустим, выбран алгоритм сравнения. Для того чтобы запланировать процедуры контроля, важно оценить количество срабатываний алгоритма в зависимости от размеров сравниваемых списков. Практика показывает, что эта зависимость имеет сложный характер. Приближенно ее можно разбить на три участка (см. рис. 1).

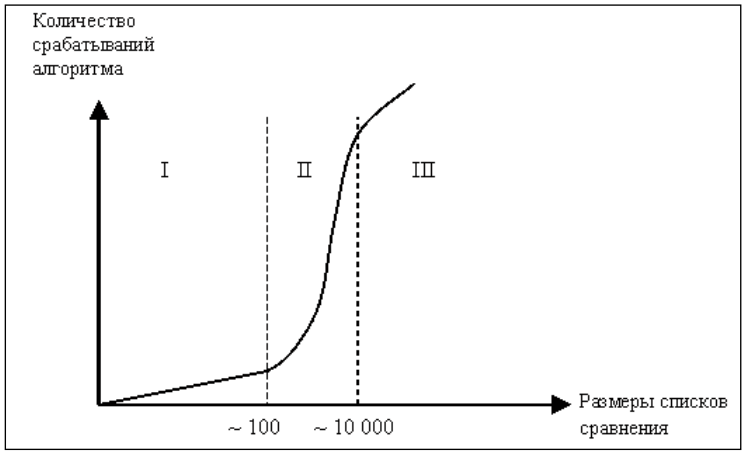

Рис. 1. Зависимость количества срабатываний алгоритма сравнения от размера списка нежелательных лиц

Пусть M — количество пассажиров на борту самолета. M не превышает величины максимальной загрузки. Размер списка нежелательных лиц — N. При N порядка нескольких сотен и меньше интересующая нас зависимость близка к линейной (на рис. 1 соответствует области I). Можно определить относительную частоту срабатываний алгоритма при сравнении одного из пассажиров с одним из пунктов списка нежелательных лиц. Математическое ожидание общего числа срабатываний алгоритма при обработке одного рейса будет определяться произведением этой относительной частоты на $M*N$ (согласно теореме о математическом ожидании числа появлений события в независимых испытаниях [5]).

При увеличении N до нескольких тысяч рассматриваемая нами зависимость становится степенной (см. область II на рис. 1). Причиной роста производной зависимости является возрастание относительной частоты срабатываний алгоритма при увеличении N. В рассматриваемой области начинает сказываться ограниченность вариативности имен людей, вследствие чего имеет место ситуация, когда при большом количестве уникальных имен в списке нежелательных лиц увеличивается вероятность того, что сбой в работе алгоритма приведет к ложному срабатыванию. При этом возникает необходимость использования лингвистического подхода.

В работе [6] проводится исследование распространенности русских фамилий. Результаты переписи показывают, что 14 000 уникальных фамилий носит большая часть переписанного населения (примерно 700 тысяч человек из миллиона). Таким образом, для списка из 10 000 человек в результате выборки, составленной произвольным образом, вероятность встретить ту или иную распространенную фамилию близка к 100 %. В мировых средствах массовой информации

периодически появляется информация о недопущении тех или иных людей на рейсы в Северную Америку или Евросоюз, в связи с тем что их имена совпадают или похожи на имена разыскиваемых лиц, причастных к террористической деятельности.

При N, имеющем порядок 10 000, нарастание производной исследуемой зависимости начинает снижаться (на рис. 1 соответствует области III). Возникает эффект насыщения относительно фактора вариативности имен людей.

Следует подчеркнуть, что границы областей, отмеченные на рис. 1, существенно зависят от национальных особенностей региона, формирующего пассажиропоток, и параметров алгоритма нечеткого поиска.

4. Оптимальные параметры алгоритма

Будем считать сравниваемые имена пассажиров и лиц нежелательного списка некоторыми текстовыми строками. Пусть выбран алгоритм поиска по сходству. Наиболее известные алгоритмы для нечеткого сравнения строк, учитывающие то, что строки могут быть разной длины, — расстояние Левенштейна и расстояние редактирования. Расстояние определяет, сколько преобразований символов необходимо сделать, чтобы преобразовать одну строку в другую. Строки считаются совпадающими, если расстояние между ними меньше заданного значения L_{max} [7. С. 192].

С увеличением длины строки возрастает вероятность искажений. Поэтому величину L_{max} алгоритма поиска по сходству, как справедливо указано в работе [8], следует брать пропорционально длине анализируемых слов, из которых состоят сравниваемые имена людей. Т. е. $L_{max} = L_0 * \rho(str1, str2)$, где L_0 — выбранная константа, а $\rho(str1, str2)$ — количество символов в более длинной из двух сравниваемых строк. Можно также рассматривать не количество символов, а количество слогов.

По мнению авторов данной статьи, наилучший результат достигается, если в качестве $\rho(str1, str2)$, т. е. допустимой меры расхождения, принять количество гласных букв (это примерно соответствует количеству слогов) и, кроме того, увеличивать эту меру для каждой последовательности из трех или большего числа подряд идущих согласных. Последовательности из нескольких идущих подряд согласных, как правило, трудно произносимы, такие последовательности могут служить обозначением букв национальных алфавитов, отсутствующих в латинском. Все перечисленное увеличивает вероятность искажения имен людей при написании с использованием латинских букв и требует ослабления критерия совпадения строк.

Выбор константы L_0 определяется размерами сравниваемых списков и уровнем процедуры контроля. Пусть осуществляется обработка списка пассажиров с возможными искажениями в написании на предмет совпадения со списком нежелательных лиц. Пусть K — количество совпадающих имен людей в случае, если осуществить самую тщательную проверку, какую только возможно. Рассмотрим данную ситуацию при различных значениях L_0 (см. рис. 2).

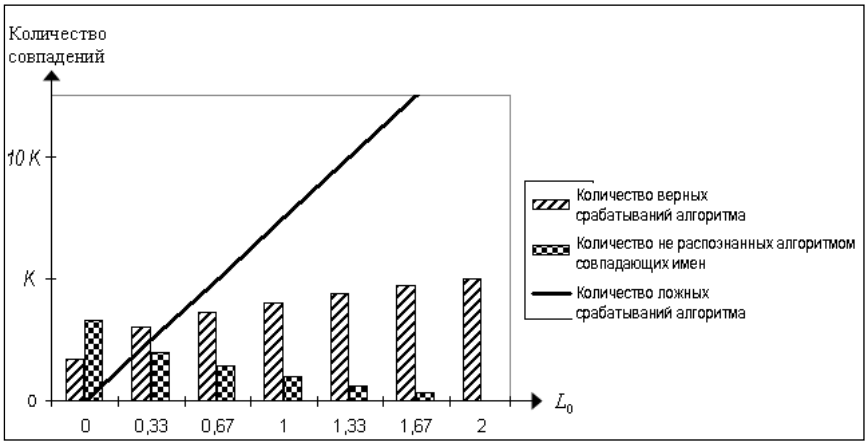

Рис. 2. Соотношения верных и ложных распознаваний при различных значениях константы L_0

Значение $L_0 = 0$ соответствует ситуации, когда имена людей признаются совпадающими, если они полностью совпадают в написании. При этом часть имен, написанных с искажениями, не распознается, но и ложных срабатываний алгоритма не будет. Ослабляя критерий совпадения строк путем увеличения значения константы L_0, мы увеличиваем долю распознанных имен, т. е. количество верных срабатываний алгоритма стремится к величине K. Практика показывает, что каждый шаг ослабления критерия совпадения строк вносит все меньший и меньший вклад в количество верно распознанных имен. Например, на рис. 2 при ($L_0 > 1$) количество нераспознанных алгоритмом совпадающих имен можно считать несущественным, если данный уровень безопасности не требует тотального выявления всех совпадений.

Важно отметить, что с увеличением значения L_0 нарастание количества ложных срабатываний алгоритма соответствует зависимости, близкой к экспоненциальной (на рис. 2 она представлена в виде прямой, при том что ось ординат имеет логарифмический масштаб). Причина такого поведения этой зависимости объясняется тем, что количество совпадающих имен в двух списках в нашем примере ограничено (равно K), с другой стороны, с ослаблением критерия совпадения строк в срабатывание алгоритма попадает все большая и большая часть рассматриваемых списков. Можно ослабить критерий настолько, что любое имя будет математически совпадать с любым другим.

Точный вид зависимостей, отображенных на рис. 2, определяется размерами сравниваемых списков и степенью искажений имен людей в них. Из этого качественного примера становится очевидным, что параметры алгоритма (в первую очередь значение L_0) требуется выбирать в соответствии с поставленной задачей контроля. Например, авиакомпании хотели бы отказывать в перевозке людям, которые ранее были отмечены в административных инцидентах на борту в связи с чрезмерным потреблением спиртных напитков. Последствия нераспознания написанного с искажениями имени лица, включенного в этот список, не являются критичными. При этом можно использовать простой алгоритм сравнения, в частности брать в качестве L_0 небольшую величину. В других случаях невыявление лица из нежелательного списка может быть критичным.

Таким образом, усиление контроля в системе безопасности вызывает непропорциональное увеличение затрат на ее модификацию и обеспечение работоспособности из-за необходимости обработки большего количество предположительных совпадений и необходимости усложнения алгоритмов. Поэтому параметры системы должны быть выбраны адекватными стоящей задаче.

5. Накопление статистической информации по работе алгоритма

Рассмотрим соотношение верных распознаваний и ложных срабатываний алгоритма в зависимости от списка нежелательных лиц. Пусть количество ложных срабатываний доминирует над правильными распознаваниями, как это показано на рис. 2 для значений величины $L_0 \geq 0{,}67$. Тогда имеет место ситуация, приведенная на рис. 3.

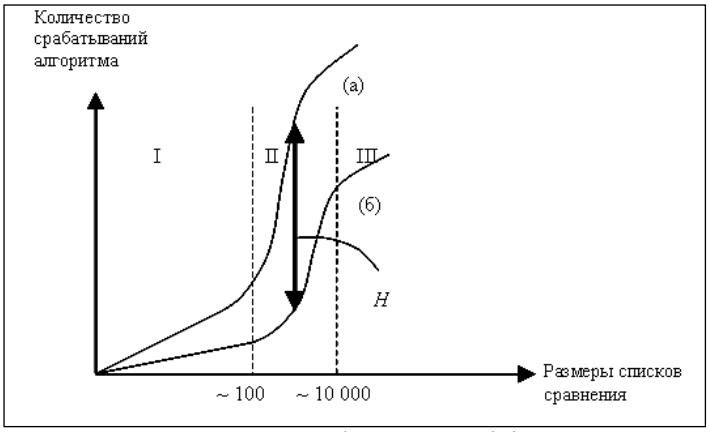

Рис. 3. Зависимости количества ложных срабатываний (а) и правильных срабатываний (б) алгоритма сравнения от размеров анализируемых списков

Обе эти зависимости от размера списка нежелательных лиц имеют одну природу и имеют ту же структуру областей I, II и III, которая была рассмотрена выше. Но поскольку кривая (а) лежит над кривой (б), то эти области у данных кривых смещены относительно друг друга.

На рис. 3 H — разброс между ложными и правильными срабатываниями алгоритма. В наихудшем случае H может достигать двух порядков. Эта ситуация возникает, когда кривая (а) лежит в области II, т. е. является степенной, а кривая (б) еще остается линейной (лежит в области I).

Здесь следует указать на необходимость накопления базы данных по правильным и ложным распознаваниям имен. Во всех предыдущих примерах описаны случаи однократной сверки пассажиров со списком нежелательных лиц с помощью алгоритма нечеткого поиска. Но после каждой проверки специалист системы контроля из выданных ему в результате работы алгоритма предположительных совпадений принимает окончательное решение о том, насколько совпадают эти имена. Очевидно, эту информацию следует запоминать, формируя базу данных имен и различных их вариантов написания с искажениями, которые встречались до этого. Еще более важно накапливать информацию о ложных срабатываниях алгоритма с целью пресекать их в дальнейшем.

При этом эффективность работы системы контроля будет определяться размерами списка нежелательных лиц и скоростью его обновления. Если список нежелательных лиц достаточно статичен, то для него быстро будет сформирована база данных ложных срабатываний и при осуществлении проверки количество ложных срабатываний алгоритма будет того же порядка или меньше количества верно распознанных имен.

Выводы

Таким образом, при создании системы контроля пассажирских списков на авиационном транспорте следует иметь в виду, что:

1. Для максимально продуктивного контроля требуется наличие базы данных ложных совпадений, описанной выше. Следует отметить, что ведение этой базы данных требует дополнительной экспертной работы специалистов, осуществляющих контроль. Эффективность использования базы данных определяется скоростью обновления списков нежелательных лиц, на предмет совпадения с которыми контролируется пассажиропоток.

2. Практика показывает, что для более уверенного контроля целесообразно использовать два подхода: поиск по сходству и лингвистический. В ситуации открытости границ и международных пассажиропотоков крайне сложно применять общие лингвистические подходы, так как граждане разных стран имеют свои индивидуальные национальные особенности написания и транслитерации в латинский шрифт. Но можно привязываться в лингвистическом оформлении к тем регионам, которым соответствует список нежелательных лиц.

3. Зависимость количества срабатываний алгоритма от размеров анализируемых списков носит сложный характер. Поэтому система контроля, построенная как на поиске по сходству, так и на лингвистическом подходе, чувствительна к объему сравниваемой информации. При увеличении списков пассажиров или нежелательных лиц может непропорционально вырасти количество срабатываний алгоритма и увеличиться соотношение между верными распознаваниями и ложными. Это существенно затрудняет планирование процедур контроля.

СПИСОК ЛИТЕРАТУРЫ:

1. *Groenewege A. D.* Compendium of International Civil Aviation. Second Edition. International Aviation Development Corporation. 1998/1999.

2. Приказ Министерства транспорта Российской Федерации (Минтранс России) от 28 июня 2007 г. № 82 г. Москва «Об утверждении Федеральных авиационных правил «Общие правила воздушных перевозок пассажиров, багажа, грузов и требования к обслуживанию пассажиров, грузоотправителей, грузополучателей».
3. *Бачурин Е. В.* Система ведения договоров с агентами как основа мониторинга продаж в авиакомпании // Известия высших учебных заведений. Северо-Кавказский регион. Серия: Технические науки. 2006. № 3. С. 103–105.
4. *Бойцов Л. М.* Поиск по сходству в документальных базах данных: хеширование по сигнатуре — оптимальное соотношение скорости поиска, простоты реализации и объема индексного файла // Программист. 2001. № 1.
5. *Гмурман В. Е.* Теория вероятностей и математическая статистика. Учебное пособие для вузов. 7-е изд., стер. М.: Высшая школа, 1999.
6. *Балановская Е. В., Соловьева Д. С., Балановский О. П., Чурносов М. И., Сорокина И. Н., Евсеева И. В., Аболмасов Н. Н., Почешхова Э. А., Серегин Ю. А., Пшеничнов А. С.* и др. «Фамильные портреты» пяти русских регионов // Медицинская генетика. 2005. Т. 4. № 1. С. 2–10.
7. *Кулябичев Ю. П., Коновалов Р. В.* Методика контроля бронирования мест пассажиров на международных рейсах авиакомпаний // Научная сессия МИФИ-2003. Т. 12. Информатика и процессы управления. Компьютерные системы и технологии.
8. *Колесов Д. А.* Разработка алгоритма поиска информации в базах данных с использованием функции нечеткого сравнения строк // Исследования по информатике. 2004. № 7. С. 125–132.