УДК 004.624

ПРОБЛЕМА ИДЕНТИФИКАЦИИ ОБЪЕКТОВ УЧЕТА В РАСПРЕДЕЛЕННЫХ СЛАБОСВЯЗАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ СБОРА ПЕРСОНАЛЬНЫХ ДАННЫХ

А. П. Толстобров[*], А. Е. Докучаев[**], В. В. Радюшкин[**], А. В. Драгунов[***]

[*] *Воронежский государственный университет*
[**] *ЗАО «РДТеХ»*
[***] *Региональный центр информатизации Псковской области*

Статья посвящена проблеме идентификации объектов учета в информационных комплексах сбора персональных данных физических лиц, особенностями которых является территориальное распределение и слабая связанность составляющих их информационных подсистем, выражающаяся в отсутствии единого управляющего комплекса, например, в виде одной общей СУБД, обеспечивающего на уровне системы в целом механизм поддержки целостности и непротиворечивости данных. Это приводит к возникновению целого ряда принципиально важных практических проблем, связанных с обеспечением необходимого уровня целостности и непротиворечивости информационного наполнения таких систем.

Одним из видов информационных систем, широко используемых и развиваемых в настоящее время, являются системы регионального или федерального уровня, осуществляющие сбор и накопление в базах данных персональных сведений, касающихся конкретных физических лиц. Это, например, такие системы как:

— базы персональных данных граждан в системе информационного обеспечения служб обязательного медицинского страхования;

— базы персональных данных граждан в службах пенсионного фонда;

— Федеральная база свидетельств (ФБС) результатов сдачи единого государственного экзамена (ЕГЭ) [1, 2, 3];

— Федеральная база персональных данных абитуриентов, реализуемая в рамках проекта по созданию Единой системы приема в вузы [1, 2, 3].

Общим для этих систем является то, что в них имеется информационная система, обслуживающая центральную — «главную» базу данных персонифицированных сведений, и большое число автономных локальных подсистем, в которых осуществляется сбор (ввод) первичных персональных данных физических лиц, и из которых данные в определенном формате периодически передаются в центральную базу данных. Таким образом, главными особенностями такого рода систем является *распределенность* и *слабая связанность* составляющих их подсистем, выражающаяся в отсутствии единого программного управляющего комплекса, например, в виде одной общей СУБД, обеспечивающего на уровне системы в целом механизм поддержки целостности и непротиворечивости данных. Это приводит к возникновению целого ряда принципиально важных практических проблем. В данной работе эти проблемы рассматриваются на примере информационной системы «Единая система приема в вузы» (ЕСП), реализуемой в рамках проекта, выполняемого по заданию Федеральной службы по надзору в сфере образования и науки, в разработке и апробации которой Воронежский государственный университет, Псковский РЦИТ и ЗАО «РДТеХ» участвовали в 2004—2005 годах [1, 2, 3].

Создание такой системы связано с расширением эксперимента по аттестации выпускников школ в форме ЕГЭ и использованием результатов этих экзаменов в качестве вступительных испытаний при поступлении в вузы. Выпускники школ, сдавшие выпускные экзамены в форме ЕГЭ, имеют теперь возможность подачи заявления для участия в конкурсах одновременно на несколько специальностей/направлений в разные вузы разных городов страны. Очевидно, что право абитуриентов одновременного участия в конкурсах на разные специальности в разных вузах и городах может быть полноценно реализовано только при обеспечении получения ими адекватной информации о

реальных конкурсных позициях в соответствующих рейтинговых списках, построенных на основании полученных абитуриентами баллов ЕГЭ, защитываемых в качестве вступительных испытаний. Создание и апробация федеральной информационной системы «Единая система приема», интегрирующей поступающую от вузов информацию о заявках абитуриентов на участие в конкурсах на поступление в разные вузы и взаимодействующую с информационными системами ФБС и РЦОИ (Региональных центров обработки информации), как раз призвано для решения задачи информирования абитуриентов на федеральном уровне.

В апробируемой системе ЕСП взаимодействие локальных вузовских подсистем с базой данных центральной системы состоит в периодической загрузке передаваемых из локальных баз массивов персональных данных в центральную базу данных. При этом в момент загрузки в центральной базе данных ЕСП уже могут присутствовать записи, относящиеся к тому же физическому лицу (абитуриенту), представляемому записью загружаемого массива данных, появившиеся там в результате предшествующего акта загрузки данных из информационных систем того же самого или другого вуза.

Принятие решения о принадлежности загружаемой персональной записи и записи, уже находящейся в базе данных ЕСП одной и той же личности осуществляется на основе сравнения значений набора ключевых полей этих записей, обеспечивающих идентификацию связанной с записью базы данных личностью.

Как показала практика, задача такой идентификации в рассматриваемых слабо связанных информационных системах на практике осложняется целым рядом объективно присутствующих факторов:

— Во-первых, первичные данные об абитуриентах (значения полей соответствующих записей) вводятся и формируются (как правило, вручную) в *разнородных* вузовских системах, поддерживающих самые разнообразные, присущие конкретному вузу, технологии организации приема и обработки информации и способы обеспечения ее достоверности.

— Первичные данные результатов сдачи ЕГЭ в базах данных РЦОИ, а затем в ФБС, формируются с использованием средств электронного сканирования и автоматизированного распознавания соответствующих полей бланков ЕГЭ.

— Данные об абитуриентах загружаются в центральную базу ЕСП *не из одного*, а из *разных*, причем не зависимых источников (при подаче абитуриентом документов в разные вузы), периодически в течение всего времени проведения приемной кампании (во время подачи абитуриентами документов, сдачи ими вступительных испытаний и, наконец, зачисления).

— В информационной системе ЕСП полностью отсутствует возможность оперативной сверки электронных данных с их первичным документальным источником (все документы абитуриентов находятся в соответствующих вузах).

— Важным моментом является то, что по результатам обработки и интерпретации накапливаемых в информационной системе ЕСП данных могут приниматься решения, имеющие в ряде случаев для абитуриентов принципиальное значение. Это делает очень высокими цены возможных ошибок и уровень ответственности при принятии таких решений.

В этих условиях главные причины сложности проблемы идентификации персональных записей абитуриентов и возможных ошибок при ее решении состоят в следующем.

1. Объективная неизбежность возникновения ошибок в значениях ключевых атрибутов записей об абитуриенте, возникающих при ручном вводе первичных данных операторами или при автоматическом распознавании бланков ответов при сдаче ЕГЭ при их обработке в РЦОИ.

2. Недостаточная идентификационная способность традиционных идентификаторов личности. Например, результаты обработки данных об абитуриентах в федеральной базе данных ЕСП показали реальность существования у *разных* абитуриентов *одинаковых* значений серии и номера паспорта (главного документа, удостоверяющего личность в нашей стране). Другими словами, приходится, к сожалению, констатировать, что, по-видимому, на данный момент среди документов, предъявляемых при поступлении в вуз, нет документов с формальным идентификатором, позволяющим *абсолютно однозначно* идентифицировать личность абитуриента. Кроме того, нельзя исключать случаи предъявления абитуриентом при подаче документов в разные вузы разных документов, удостоверяющих его личность.

3. Наконец, важным моментом, существенно влияющим на сложность решения задачи,

является очень большой объем базы данных ЕСП в силу ее федерального характера.

Именно сочетание этих факторов на практике порождает большую неопределенность и сложности в поисках приемлемых решений проблемы персональной идентификации персональных записей.

Обычно при сопоставлении двух персональных записей для идентификации связанных с ними личностей используется двухальтернативная процедура принятия решения — {«записи представляют *одну* личность», «записи представляют *разные* личности»}. Решение принимается по совпадению или несовпадению значений выбранного набора ключевых полей записи. При этом возможны ошибки идентификации двух следующих видов.

Ошибка первого рода — *отсутствие идентификации вопреки ее фактическому наличию*. Идентификация представляющих записи личностей *не произошла*, не смотря на их фактическую принадлежность одной и той же личности (ключевые поля записей, принадлежащих одной и той же личности, не совпали, например, из-за ошибок при вводе значений этих полей).

Ошибка второго рода — *ложная идентификация записей, принадлежащих двум различным личностям*. У записей, принадлежащих разным личностям, по тем или иным причинам совпали значения ключевых полей, по которым производится идентификация записей.

Рассмотрим более подробно *ситуации*, которые могут возникнуть при идентификации записей при загрузке данных в центральную базу данных ЕСП с учетом изложенных выше обстоятельств.

Первая ситуация — при идентификации произошла *ошибка первого рода* — хотя на самом деле записи принадлежат одному абитуриенту, в базе данных они присутствуют как записи о разных абитуриентах. Такая ошибка может произойти по следующим двум причинам.

Причина 1 — несовпадение ключевых атрибутов записей из-за *ошибок оператора* при вводе первичных данных абитуриента (значений этих атрибутов).

— *Вероятность* появления такой ошибки зависит от используемой технологии ввода первичных данных и увеличивается при увеличении количества ключевых полей и при наличии среди них полей с текстовыми значениями, ввод которых, как правило, осуществляется вручную.

— *Последствие и опасность* ошибки. Если ошибка не будет обнаружена, в базе данных окажется две записи, относящихся к одной и той же личности. Загруженная запись с ошибочными значениями в ключевых полях окажется для пользователей *не видимой*, так как при запросах с указанием правильного значения ключевых полей эта запись не будет найдена.

— *Предотвращение ошибки* при загрузке данных в центральную базу данных невозможно, так как ошибки при вводе первичных данных совершаются на местах, в вузовских системах.

— Возможности обнаружения ошибки:

1) проведение последующего анализа записей базы данных с использованием более широкого набора идентификационных признаков персоны и более сложного алгоритма сопоставления значений сравниваемых полей, в принципе, позволяет выявить записи в той или иной степени *подозрительные* с точки зрения их возможной принадлежности одной личности;

2) ошибка может быть обнаружена самим абитуриентом, когда при обращении с запросом к базе данных, он не обнаруживает своей персональной информации в ответе на свой запрос.

Причина 2 — абитуриент умышленно или не умышленно в двух актах подачи документов заявил разные значения ключевых атрибутов, например, при подаче документов в одном вузе он предъявил данные одного паспорта, а в другом данные другого паспорта (потерял паспорт, украли, спрятал, ...), изменил фамилию, и т.д. (практика показывает, что такие ситуации являются вполне реальными).

— Сознательный умысел таких действий абитуриента можно объяснить, возможно, его попыткой скрыть результаты школьного этапа ЕГЭ или желанием скрыть факт подачи документов в разные вузы. При действительной утере документа, удостоверяющего личность, по идее, абитуриент заинтересован в сообщении этого факта в приемную комиссию.

— *Последствие ошибки* — в базе данных окажется две записи, относящихся к одной и той же персоне.

— *Опасность ошибки* состоит в возможности злоупотреблений со стороны абитуриента,

например, в случае его попытки сокрытия результатов сдачи ЕГЭ.

— *Возможность обнаружения ошибки.* В принципе возможно выявление подозрительных на принадлежность одному лицу записей при проведении дополнительного анализа по расширенному набору полей, позволяющих идентифицировать личность (например, кроме серии-номера паспорта можно сравнивать номер документа об образовании и др.). Наложение на эту ситуацию ошибок в ключевых полях из-за некорректного ввода их значений существенно усложняет такой анализ.

Рассмотрим вторую ситуацию, когда произошла ошибка второго типа (*ложная идентификация*) — две записи, фактически принадлежащие *разным* абитуриентам, в результате совпадения ключевых полей интерпретируются, как принадлежащие одному лицу. Такая ошибка может произойти по следующим причинам.

Совпадение ключевых полей для записей, принадлежащих разным персонам из-за недостаточной идентификационной способности выбранного набора ключевых атрибутов (совпали серия и номер паспорта для разных персон) и совпадение ключевых полей для записей, принадлежащих разным персонам из-за ошибок при вводе значений ключевых атрибутов (в результате чего произошло их случайное совпадение).

— *Вероятность появления такой ошибки* уменьшается при увеличении количества ключевых полей, по сравнению которых принимается решение.

— *Последствие ошибки* — в базе ЕСП будут объединены данные записей, относящихся к разным персонам, или данные записи одной персоны заместят данные записи другой персоны, что приведет к *невосстановимому* искажению или потере данных, делая высокой цену такой ошибки.

— *Предотвращение ошибки* возможно при соответствующем выборе набора ключевых полей. В связи с тем, что, как показала практика, в нашем распоряжении нет одного такого надежного идентификатора личности, ключ должен быть *составным* и включать в себя *набор* полей, например, помимо серии и номера паспорта, поля ФИО, номер документа об образовании или какие-либо еще. Однако при возможном наличии в загружаемых данных ошибок ввода значений ключевых полей, такое решение будет серьезно осложнять решение задачи обнаружения записей, принадлежащих одному абитуриенту, увеличивая вероятность ошибок первого рода. Очевидно, что увеличение числа полей в ключе и наличие в нем текстовых полей, вводимых вручную, будет приводить к увеличению вероятности их несовпадения из-за возможных ошибок при их вводе. При этом возможны две отличающиеся друг от друга ситуации.

1. Вузом повторно передается в ЕСП запись об абитуриенте для того, чтобы исправить или дополнить загруженные ранее данные об абитуриенте. В этом случае идентификация записей может осуществляться по регистрационному номеру абитуриента из вузовской информационной системы.

2. Новая запись об абитуриенте передается из информационной системы другого вуза, в который также поступает тот же самый абитуриент. В этом случае идентификация записей абитуриентов осуществляется по ключевым полям, значения которых вводились в *разных* системах. Вследствие этого существует вероятность несовпадения значений ключевых полей из-за ошибок при их вводе в разных вузах.

Таким образом, задача выбора набора ключевых полей при наличии ошибок в идентифицируемых записях наталкивается на противоречивые требования. Рассмотренные два вида ошибок не являются независимыми. Меры по минимизации воздействия для одного типа ошибки, приводят к увеличению ущерба от другого вида ошибок, в частности, увеличение количества ключевых полей снижает вероятность ошибки второго рода, то есть *ложной идентификации*, но повышает вероятность ошибки первого рода — *отсутствия идентификации*.

Представляется, что последствия ложной идентификации записей, принадлежащих двум разным персонам (ошибки второго рода), являются гораздо более опасными, чем последствия ошибки не идентификации двух записей для одной персоны (ошибки первого рода), так как даже при последующем обнаружении такой ошибки ее следы в базе данных устранить сложнее из-за необратимого искажения данных. Поэтому при выборе состава полей, по сравнению значений которых принимается решение, следует исходить в первую очередь из того, чтобы минимизировать возможность появления

именно таких ошибок. Этот фактор мотивирует увеличение числа ключевых полей, по совпадению которых принимается решение о принадлежности записей одной персоне.

Приведенные рассуждения позволяют сделать вывод, что в рамках двухальтернативной процедуры принятия решений по идентификации персональных записей в реальных условиях невозможно обеспечить приемлемую надежность обеспечения адекватности информации в центральной базе персональных данных при ее пополнении путем загрузки данных, поступающих из вузовских информационных систем.

Решение проблемы, в гораздо большей степени удовлетворяющее необходимым требованиям, можно получить при расширении набора ключевых полей и использовании процедуры принятия решения, в которой, кроме принимаемых с необходимой степенью уверенности решений вида — «записи представляют *одну* личность» или «записи представляют *разные* личности», в случаях, когда эта уверенность оказывается недостаточной, принимается решение о проведении дополнительного, более детального анализа значений полей сравниваемых персональных записей.

Пусть, например, набор ключевых полей, по которым осуществляется идентификация записей, включает в себя четыре поля. Для записи A — $\{A_1, A_2, A_3, A_4\}$ и для записи B — $\{B_1, B_2, B_3, B_4\}$.

Решение о том, что идентифицируемые записи представляют *одну и ту же* личность принимается в случае, когда выполняется условие

$$A_1 = B_1, A_2 = B_2, A_3 = B_3, A_4 = B_4.$$

Решение о том, что идентифицируемые записи представляют *разные* личности принимается в случае, когда выполняется условие

$$A_1 \neq B_1, A_2 \neq B_2, A_3 \neq B_3, A_4 \neq B_4.$$

В ситуации, возникающей в случае, когда совпадение имеет место для одного или нескольких, но не всех, ключевых полей (например, $A_1 = B_1, A_2 = B_2, A_3 \neq B_3, A_4 \neq B_4$), записи относят к третьей группе записей «подозрительных» с точки зрения их принадлежности одной личности и для них должны применяться дополнительный анализ значений их полей и более сложная процедура принятия решения. Например, у двух записей значения поля **серия_номер_паспорта** совпали, а поля **дата_рождения** НЕ совпали, или значения полей **серия_номер_паспорта**, **дата_рождения** и **фамилия** совпали, а значение поля имя не совпало.

Очевидно, что само выделение такой категории не идентичных, но «похожих» записей, «подозрительных» с точки зрения возможного представления ими одной и той же личности, в большой степени обусловлено фактом возможного искажения значений полей из-за ошибок ввода первичных данных. При абсолютно надежных значениях полей при любых несовпадениях записи следовало бы относить к разным личностям. Учитывая то, что многие описывающие личность поля обладают определенной семантикой, в ряде случаев возможно обнаружение факта отличия записей, принадлежащих одной личности, из-за ошибок ввода первичных данных. Совершенно естественно, например, отнести к одной личности записи, у которых совпадают значения всех ключевых полей кроме значения поля **Имя**, в котором значения отличаются всего лишьодним символом.

В случаях, когда анализ сравниваемых записей не позволяет принять надежного решения по идентификации персональных записей, обычно формируется протокол с указанием на возникшую проблему. В силу того, что при загрузке в базу данных ЕСП отсутствует возможность оперативного доступа к первичным документам для верификации значений сомнительных полей в базе данных по этим документам (все документы находятся на местах в приемных комиссиях вузов), такая верификация сомнительных значений может осуществляться только путем отправки соответствующих сообщений в вузы, из которых поступили эти данные.

Изложенные выше соображения позволили выработать процедуру идентификации персональных записей абитуриентов при загрузке данных, получаемых из вузовских систем, в центральную базу данных ЕСП, которая была реализована и практически апробирована в 2005 году.

Эта процедура выполняется периодически при получении из конкретной вузовской информационной системы массива данных для догрузки этих данных в центральную базу данных системы. При ее выполнении каждая запись массива данных, получаемого из информационной системы вуза, сверяется с каждой записью главной базы данных. В качестве полей, которые используются для идентификации личности абитуриента, представляемого участ-

вующими в процедуре сравнения записей, используются значения следующих полей или групп полей:

1) регистрационный номер (идентификатор) абитуриента в информационной системе вуза и идентификатор вуза;

2) фамилия, имя, отчество и дата рождения абитуриента;

3) группа полей, идентифицирующих документ, удостоверяющий личность (тип документа, серия, номер, дата выдачи);

4) группа полей, идентифицирующих документ об образовании (тип документа, серия, номер, дата выдачи).

Каждой их этих групп полей ставится в соответствие один из четырех флагов:

1) Рег № в вузе (регистрационный номер абитуриента в вузе),

2) **ФИО** и **ДР** (фамилия, имя, отчество и дата рождения),

3) **Уд Док** (документ, удостоверяющий личность),

4) **Док Обр** (документ об образовании).

По результату сравнения значений перечисленных групп полей каждой пары сопоставляемых записей эти флаги устанавливаются в соответствующие состояния, указывающие на наличие совпадения или несовпадения значений в соответствующей группе полей. Последующая обработка этих состояний и принимаемые решения состоят в следующем.

Если все четыре флага показывают *несовпадение* значений во всех сравниваемых группах ключевых полей, то принимается решение, что две сравниваемые записи принадлежат разным абитуриентам. Если получаемая из вузовской системы запись оказывается не проидентифицированной ни с одной записью центральной базы данных, то принимается решение, что обрабатываемая запись полученного от вуза массива данных соответствует новому абитуриенту и производится ее добавление (загрузка) в центральную базу данных системы.

Если все четыре флага указывают на полной совпадение значений сравниваемых полей, то, принимается решение, что обе записи принадлежат одному и тому же абитуриенту. В этом случае выполняются действия по коррекции (замене на новые) значений полей в записи центральной базы, не участвующих в процедуре идентификации, например, полей — **пол**, **код страны**, **признак медалиста**, **код типа образования** и т.д.

Рассмотрим ситуации, когда при совпадении значений полей флага **Рег №** в вузе один или несколько остальных флагов указывают на *не*совпадение значений соответствующих полей. Здесь важно отметить, что вполне оправданно считать, что значения полей, соответствующих флагу **Рег №** в вузе, не могут быть искаженными, так как они наверняка формируется в информационной системе вуза при вводе данных нового абитуриента автоматически. Напротив, значения других полей, как правило, вводятся оператором *вручную* и, вследствие этого, приходится принимать во внимание возможность того, что значения этих полей могут искажаться и, поэтому отличаться для одного и того же абитуриента из-за ошибок (опечаток) при вводе данных. Практика говорит о том, что такого рода ошибки, как правило, приводят к искажениям представляющих значения полей символьных строк, выражающимся в несовпадении небольшого числа символов, чаще всего одного символа.

Для проверки предположения о том, что причиной несовпадения значений сравниваемых полей, на самом деле представляющих одного и того же абитуриента, являются ошибки ввода, производится сравнение значений «подозрительных» полей по более сложному алгоритму, учитывающему лингвистическую близость (похожесть) сравниваемых значений полей. Для определения лингвистической похожести значений полей был использован так называемый алгоритм Левенштейна [4], в котором мерой разницы двух последовательностей символов (строк) является минимальное количество операций вставки, удаления и замены, необходимых для перевода одной строки в другую. Таким образом, если строки полностью совпадают, то дистанция Левенштейна равна нулю, если они отличаются одним символом, то единице, и т.д. Будем называть «слабыми» отличия, для которых дистанция Левенштейна не превышает заданного значения. В рассматриваемом, практически реализованном варианте алгоритма сравнения значения полей считаются «похожими», если они различаются не более, чем в одном символе (мера различия Левенштейна не более 1).

Учитывая сказанное, далее в описываемом алгоритме идентификации решения вырабатываются следующим образом.

1. Если поля флага **Рег №** в вузе совпадают, и, кроме того, при полном совпадении полей

остальных флагов имеют место «слабые» в смысле дистанции Левенштейна отличия:

— только значений полей флага **ФИО** и **ДР** (фамилия, имя, отчество и дата рождения);

— или только значений полей флага **Уд Док** (тип/серия/номер/дата выдачи документа, удостоверяющего личность);

— или только значений полей флага **Док Обр** (тип/серия/номер/дата выдачи документа об образовании);

— или значений полей обоих документов **Уд Док** и **Док Обр** одновременно;

— или одновременно значений полей флага **ФИО** и **ДР** и **Уд Док**;

— или одновременно значений полей флага **ФИО** и **ДР** и **Док Обр**;

то значения полей в базе данных корректируются (заменяются на новые).

2. Если имеют место «слабые» отличия в группах полей всех трех флагов **ФИО** и **ДР**, **Уд Док** и **Док Обр**, при полном совпадении полей флага **Рег №** в вузе, то формируется и записывается в журнал ошибок сообщение об ошибке вида: «*Под регистрационным номером XXX были ранее занесены данные об абитуриенте ФИО, ДАТА РОЖДЕНИЯ, с другим документом, удостоверяющим личность и другим документом об образовании. Удалите абитуриента с регистрационным номером XXX, измените регистрационный номер и повторно загрузите документ, или же исправьте ошибки в ФИО и документах*», которое отправляется в вуз, источник информации.

3. Если отличаются (но «слабо») только поля флага **Рег №** в вузе при полном совпадении полей остальных флагов, то:

— если записи из одного вуза (идентификаторы вузов совпадают), то принимается решение, что они принадлежат одному абитуриенту и производится обновление значений записи, то есть замена старого значения регистрационного номера абитуриента и значений неключевых полей на новые;

— если записи из разных вузов (не совпадают идентификаторы вузов), то принимается решение, что получены данные о заявлении того же абитуриента, но в другой вуз и эти новые данные также добавляются к уже имеющимся.

4. Если различаются значения полей флагов **Рег №** в вузе и **ФИО** и **ДР**, при полном совпадении полей флагов **Уд Док** и **Док Обр**, то принимается решение о добавлении этой записи, как записи *нового* абитуриента из другого вуза, но уведомлением в журнал ошибок о потенциальном конфликте: «*С точно такими же документами уже зарегистрирован абитуриент ПОХОЖИЙ_АБИТУРИЕНТ. (Зарегистрирован ВУЗами: СПИСОК_ВУЗОВ). В случае, если это один абитуриент, необходимо слияние.*»

5. Если различаются значения полей флагов **Рег №** в вузе и **Уд Док**, при полном совпадении полей флагов **ФИО** и **ДР** и **Док Обр**, то принимается решение о добавлении этой записи, как записи *нового* абитуриента из другого вуза, но уведомлением в журнал ошибок о потенциальном конфликте: «*С точно такими же ФИО и датой рождения и документом об образовании уже зарегистрирован абитуриент ПОХОЖИЙ_АБИТУРИЕНТ. (Зарегистрирован ВУЗами: СПИСОК_ВУЗОВ). В случае, если это один абитуриент, необходимо слияние.*»

6. Если различаются значения полей флагов **Рег №** в вузе и **Док Обр**, при полном совпадении полей флагов **ФИО** и **ДР** и **Уд Док**, то принимается решение о добавлении этой записи, как записи *нового* абитуриента из другого вуза, но уведомлением в журнал ошибок о потенциальном конфликте: «*С точно такими же ФИО и датой рождения и документом, удостоверяющим личность уже зарегистрирован абитуриент ПОХОЖИЙ_АБИТУРИЕНТ. (Зарегистрирован ВУЗами: СПИСОК_ВУЗОВ). В случае, если это один абитуриент, необходимо слияние.*»

7. Если совпадают только поля флага **Док Обр**, то такая запись в базу данных *не заносится*, и все ее данные записываются в журнал ошибок с уведомлением о конфликте: «*С точно таким же документом об образовании уже зарегистрирован абитуриент ДАННЫЕ_АБИТУРИЕНТА. (Зарегистрирован ВУЗами: СПИСОК_ВУЗОВ).*»

8. Если совпадают только поля флага **Уд Док**, то такая запись в базу данных *не заносится* и все ее данные записываются в журнал ошибок с уведомлением о конфликте: «*С точно таким же документом, удостоверяющим личность, уже зарегистрирован абитуриент ДАННЫЕ_АБИТУРИЕНТА. (Зарегистрирован ВУЗами: СПИСОК_ВУЗОВ).*»

Далее рассматриваются ситуации, когда имеют место «сильные» в смысле дистанции Левенштейна различия сравниваемых записей.

9. Если все ключевые поля совпадают, но различается, причем «сильно», только документ об образовании (флаг **Док Обр**), то принимается решение, что абитуриент тот же самый, но это, возможно другой документ. В этом случае исправляются данные о документе, если он того же типа, что и в старой записи, или добавляется новый документ, если его тип отличается.

10. Если все ключевые поля совпадают, но различается, причем «сильно», только документ, удостоверяющий личность (флаг **Уд Док**), то принимается решение, что абитуриент тот же самый, но это, возможно другой документ. В этом случае также исправляются данные о документе, если он того же типа, что и в старой записи, или добавляется новый документ, если его тип отличается.

Как уже говорилось, в практически реализованной процедуре идентификации представляемых записями абитуриентов при принятии решения о «похожести» записей использовалось значение дистанции Левенштейна, равное нулю или единице. Очевидно, что при использовании в качестве порогового значения этой дистанции равной двум или больше, количество «похожих» записей будет больше. Т.е. для большего числа записей будет приниматься решение о том, что они принадлежат одному абитуриенту, а значения их «слабо» отличающихся полей отличаются из-за ошибок ввода. Как следует из изложенного выше, в этом случае для большего числа записей будет приниматься решение о «слиянии» представляемых ими данных.

Увеличение порогового значения «похожести» значений полей, однако, в общем случае увеличивает вероятность того, что в категорию «похожих» и из-за этого объединяемых записей попадут запис,и на самом деле принадлежащие *разным* абитуриентам. Очевидно, что такого рода ошибка крайне нежелательны, в том числе и из-за того, что их последствия трудно устранить, даже в случае их последующего обнаружения. Это объясняет осторожность, с которой следует подходить к увеличению порогового значения «похожести» значений сравниваемых полей.

Следует обратить внимание на еще одну особенность используемого алгоритма. Как видно из вышеизложенного, в процедуре сравнения участвуют две записи — запись из загружаемого массива по очереди сравнивается с записями центральной базы данных. При этом не принимается во внимание тот факт, что, когда говорится о «похожести» записей, реально запись загружаемого массива может оказаться «похожей» не на одну, а на несколько записей центральной базы данных. Причем степень этой «похожести» может быть разной. В используемом же алгоритме эта возможность никак не учитывается, и анализ заканчивается при обнаружении первой «похожей» записи.

Все изложенное выше делает также исключительно важным рассмотрение мер, которые могли бы хоть и не исключить совсем появление неоднозначно идентифицируемых записей, но, по крайней мере, минимизировать их количество. И в этом аспекте особое значение приобретает обсуждение и выработка способов и мер радикального снижения ошибок при вводе первичных данных в системах вузовского уровня.

Стремление уменьшения количества ошибок при вводе первичных данных явилось, в частности, одной из причин, побудившей в Воронежском госуниверситете изменить в 2004—2005 годах процедуру и технологию оформления заявлений абитуриентов и ввода их содержимого в базу данных системы «Абитуриент». Если раньше использовалась общепринятая процедура, когда информация в базу данных вводилась операторами с бланка заявления, заполненного от руки абитуриентом, то в настоящее время данные заявления абитуриентов вводятся в базу данных с помощью специальной web-формы непосредственно из первичных документов самим абитуриентом или оператором в присутствии абитуриента, и уже после этого по введенным данным системой формируется печатная форма заявления. После проверки содержания распечатанного заявления абитуриентом заявление вместе с другими документами представляется в приемную комиссию, где дополнительно проверяется персоналом приемной комиссии с использованием, при необходимости, представленных абитуриентом документов. Такая процедура, устранившая потенциально опасный с точки зрения возможных ошибок этап предварительной фиксации необходимых данный в рукописном виде на бумажном носителе, и обеспечивающая, фактически, несколько рубежей контроля правильности введенных данных, позволила существенно уменьшить количество ошибок ввода.

Для дополнительного контроля правильности значений ключевых полей при их вводе в базу

данных могут использоваться дополнительные меры, например, такой механизм, как повторный ввод для подтверждения данных, подобно тому, как это делается при вводе и подтверждении пароля для входа в систему.

И, наконец, следует отметить еще один, показавший на практике свою эффективность, механизм дополнительного контроля персональных данных. Это возможность их контроля самим абитуриентом, реализованную благодаря возможности доступа абитуриента к своему электронному личному делу через web-сайт и наличие обратной связи через его форум. Такая возможность реализована, например на сайте ЕСП (http://ekp.edu-soft.ru), сайте ВГУ «Абитуриент OnLine» (http://www.abitur.vsu.ru). Бесспорно, что абитуриент, как правило, является одним из самых заинтересованных лиц в обеспечении достоверности фигурирующих в его личном деле персональных данных.

Особенности и пути решения проблемы, которой посвящена данная работа, рассмотрены на примере реализации конкретной информационной системы федерального уровня, предназначенной для информационного обеспечения абитуриентов. Однако, обсуждаемые обстоятельства имеют место и в других аналогичных системах, которые можно отнести к классу распределенных слабосвязанных информационных систем с базами данных, предназначенных для сбора персонифицированных данных физических лиц. В частности, это уже упомянутые выше системы информационного обеспечения служб обязательного медицинского страхования, базы персональных данных граждан в службах пенсионного фонда и целый ряд других аналогичных систем. Поэтому рассмотренные проблемы и пути их решения могут представлять интерес и оказаться полезными при практической реализации этих систем для повышения качества их функционирования.

СПИСОК ЛИТЕРАТУРЫ

1. *Драгунов А.В.* Интеграция информационных систем в управлении образованием региона и общероссийская система оценки качества образования. / Сборник трудов Всероссийской конференции Интеграция информационных систем в управлении образованием 2005. 24—26 марта 2005 г., Псков. С. 13—17.

2. *Толстобров А.П.* Информационные технологии в управлении образовательным учреждением. Единая система приема и зачисления в вузы. / Сборник трудов Всероссийской конференции Интеграция информационных систем в управлении образованием 2005. 24—26 марта 2005 г., Псков. С. 112—116.

3. *Радюшкин В.В.* Интеграция систем федерального и интегрального уровней в рамках федеральной системы единого конкурсного приема. / Сборник трудов Всероссийской конференции Интеграция информационных систем в управлении образованием 2005. 24—26 марта 2005 г., Псков. С. 122—126.

4. Расстояние Левенштейна. http://ru.wikipedia.org/wiki/Дистанция_Левенштейна.