МЕТОДЫ СТАНДАРТИЗАЦИИ И КЛАССИФИКАЦИИ ЗАПИСЕЙ О МЕСТЕ РОЖДЕНИЯ И ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ В ДАННЫХ ПЕРЕПИСИ ВЕЛИКОБРИТАНИИ 1851–1911 гг.

METHODS OF STANDARDISING AND CODING BIRTHPLACE STRINGS AND OCCUPATIONAL TITLES IN THE BRITISH CENSUSES 1851–1911

Шурер Кевин,
профессор, проректор по науке,
Университет Лестера
E-mail: k.schurer@leicester.ac.uk

Kevin Schürer

Пенькова Татьяна Геннадьевна,
кандидат технических наук, старший
научный сотрудник Института
вычислительного моделирования СО РАН
E-mail: penkovatg@gmail.com

Tatyana G. Penkova

Представлены методы стандартизации и классификации текстуальных записей о месте рождения и профессиональной деятельности, полученных по данным переписи населения Англии, Уэльса и Шотландии за 1851–1911 гг. Разработаны алгоритмы формирования классификационных кодов профессий и идентификации географических районов на основе сопоставления исходных и справочных данных. Предложенные методы основаны на интеграции вычислительных технологий, математических методов и экспертных знаний.

Ключевые слова: стандартизация, классификация, обработка текстуальных данных, перепись населения, Великобритания.

This paper presents a technique of standardising and coding textual birthplace and occupation strings in the censuses of England and Wales and Scotland, 1851–1911. The approaches are based upon the integration of the computer technologies, mathematical methods and expert knowledge. The classification of occupations is defined by two algorithms based on statistical decision theory in order to allocate codes from the original occupation strings. The method of standardising parishes is based on the comparison of original birthplace strings and reference data.

Keywords: Historic census data, Birthplace standardisation, Occupation coding.

Введение

Данные переписи населения — один из важнейших источников изучения широкого круга исторических вопросов, особенно связанных с социально-экономическими процессами развития территорий[1]. Несмотря на ценность данных, полученных в результате переписи, существует серьезная проблема — данные представляют собой текстовые записи (ответы респондентов), и прежде чем они будут проанализированы и интерпретированы, требуется их предварительная обработка. Данная проблема наглядно иллюстрируется результатами десятилетних переписей, проведенных в Великобритании с 1851 по 1911 г. Материал, полученный из этих переписей, представляет собой основу настоящей работы, выполненной в рамках крупного международного проекта по обработке данных переписи викторианской эпохи[2].

В общей сложности объем данных составляет 183 470 969 персональных записей. На основе этих объединенных данных сформировано 7 304 708 уникальных записей с описанием профессиональной деятельности (*Occupation*) и 6 703 779 уникальных записей с описанием места рождения (*Birthplace*). Такое количество записей объясняется тем, что ответы респондентов были очень разнообразны. Например, в записях о профессиональной деятельности профессия «WATCHMAKER» (часовщик) выражена следующими описаниями:

WATCH MACKER
WATCH MAKER & GREEN GROCER
WATCH MAKER (REPAIRER)
WATCH MAKER EMPLOYING 4 MEN & 3 BOYS
WATCH MAKER IN ALL BRANCHES
WATCH MAKER IN GENERAL
WATCH REPAIRER & MAKER
WATCH- MAKER
WATCHMAAKER
WATCHMAER
WATCHMAK
WATCHMAKER (MAS)
WATCHMAKER CLOCK
WATCHMAKER EMP 1 ASSIST 1 APPRENT
WATCHMAKER EMPLOYS 2 FEMALES & 2 MALE APP
WATCHMAKER ETC
WATCHMAKER GENERAL
WATCHMAKER MASTER EMP 1 MAN + 1 BOY
WATCHMAKER MASTER EMPLOY 1 MAN
WATCHMAKER OUT OF EMPLOY
WATCHMAKER REPAIR
WATCHMAKER SPRINGER
WATCHMAKR
WATCHMALER
WATCHMEKER
WATCHMENDER
WATCHV MAKER
WATCHWORK
WATCJMAKER
WATCKMAKER
WATCMAKER
WATHCHMAKER
WATHCMAKER
WATHMAKER
WORKING WATCHMAKER
WORKING WATCHMAKER AND JEWELLER
WQTCHMAAKER
WTAHCMAKER
WTCHMAKER
WTCHMKR
WWATCH MAKER
WWATCHMAKER

Аналогично в записях о месте рождения, например, населенный пункт «HUSBAND'S BOSWORTH», расположенный в графстве «LEICESTERSHIRE», представлен следующими вариантами:

H BOSWORTH| [BLANK] | [BLANK]
H. BOSTH|LEICESTER|ENGLAND
HBDS BOSWORTH|LEICESTERSHIRE| [BLANK]
HDS BOSWORTH LEICESTER|LEICESTERSHIRE| [BLANK]
HUSBANDS BODWORTH|LEICESTERSHIRE| [BLANK]
HUSBANDS BOSWORTH| [BLANK] |LEISTER
HUSBANDS BOSWORTH| [BLANK] | [BLANK]
HUSBANDS BOSWOR|LEICESTERSHIRE| [BLANK]
HUSBANDS BOSWTH|LEICESTER|ENGLAND
HUSBANDS BSWORTH|LEICESTERSHIRE| [BLANK]
HUSBANDS B|LEICESTER|ENGLAND
HUSBANK BOSWORTH LESTERS|LEICESTERSHIRE| [BLANK]
HUSBANS B|LEICESTERSHIRE| [BLANK]
HUSBARDS BOSWORHTH|LEICESTERSHIRE| [BLANK]
HUSBARDS BOSWORTH|LEICESTERSHIRE| [BLANK]
HUSBORNE BORWORTH|LEICESTERSHIRE| [BLANK]
HUSBOS BOSWORTH|LEICESTER|ENGLAND
HUSDS BOSWORTH|LEICESTER|ENGLAND
HUSLANDS HOSWORTH|LEICESTERSHIRE| [BLANK]
LEIC HUSBANDS BOSWORTH| [BLANK] | [BLANK]
LEICESTER HUSBOND BOSWORTH|LEICESTERSHIRE| [BLANK]
LEISTER HUSBAND BOSWORTH|LEICESTERSHIRE| [BLANK]
LENTER HUSBANDS BOSWORTH| [BLANK] | [BLANK]
LESTER HBS BOSWORTH| [BLANK] | [BLANK]
LESTER HUSBAND BASWORTH|LEICESTERSHIRE| [BLANK]
LESTER HUSBANDS BOSWORTH|LEICESTERSHIRE| [BLANK]
LESTER HUSBANDS BOSWORTH| [BLANK] | [BLANK]
LESTERSHIRE HDS BOSWORTH| [BLANK] | [BLANK]
LESTERSHIRE HUSBANDS BOSWORTH|LEICESTERSHIRE| [BLANK]
LEUICESTERSHIRE HUSBANDS RESIDENT BOSWOTH|LEICESTERSHIRE| [BLANK]
LICESTER HUSBANDS BOSWORTH| [BLANK] | [BLANK]
LIECESTERSHIRE HUSBANDS BOSWORTH|LEICESTERSHIRE| [BLANK]
LTDS BOSWORTH|LEICESTERSHIRE| [BLANK]
LUSHAND BAWSWORTH|LEICESTERSHIRE| [BLANK]
MUSBANA BOSWORTH|NORTHAMPTONSHIRE| [BLANK]

PARISH HUSBAND BOSWORTH|LEICESTERSHIRE| [BLANK]
PARISH HUSBANDS BOSWORTH|LEICESTERSHIRE| [BLANK]
RUSBAND BASWORTH|LEICESTERSHIRE| [BLANK]
RUSBUNDS BASWORTH|LEICESTERSHIRE| [BLANK]
[BLANK] |LEICESTERSHIRE|HASBORD BOSWORTH
[BLANK] |LEICESTERSHIRE|HUSBANDS BOSWORTH LODGE

Предварительный анализ исходных данных показал, что 77,7% записей о профессиональной деятельности и 70,2% записей о месте рождения имеют частоту, равную 1. Такая многовариантность свидетельствует о весьма своеобразном характере записей, при этом большой объем данных исключает возможность их ручной обработки. Следовательно, актуальной задачей становится разработка специализированных средств автоматической (полуавтоматической) стандартизации и классификации текстуальных записей переписи.

1. МЕТОД КЛАССИФИКАЦИИ ЗАПИСЕЙ О ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Метод разработан на основе технологии структурного анализа и проектирования (SADT, Structured Analysis and Design Technique)[3]. На основе SADT-методологии разработана функциональная модель процессов классификации записей о профессиональной деятельности. Модель отображает основные этапы, механизмы, посредством которых выполняются основные функции, данные, изменяемые и появляющиеся в результате выполнения функций, а также правила и ограничения выполнения функций. На рисунке 1 представлена контекстная диаграмма IDEF0 процесса классификации.

Процесс классификации записей включает три основных этапа:

О1 — создание классификаторов и справочников;

О2 — определение классификационных кодов;

О3 — оценивание результатов классификации.

На первом этапе (*О1*) осуществляется разработка и модификация необходимых справочников и классификаторов. Данный этап выполняется экспертами на основе существующих классификаторов, исторических и исходных данных.

Для указанной задачи классификации используются ранее разработанный классификатор профессий 1881 г., содержащий 1 400 000 записей, классификационные коды, разработанные для данных переписи Англии и Уэльса 1911 г., а также справочник, созданный по результатам экспертной обработки объединенных данных[4].

На втором этапе (*О2*) определяются классификационные коды. Этот этап реализуется программой на основе разработанных методов и алгоритмов с использованием созданных на предыдущем этапе справочников и классификаторов.

На третьем этапе (О3) выполняется оценивание результатов классификации с внесением необходимых корректировок в справочники и алгоритмы, формирование итоговых результатов.

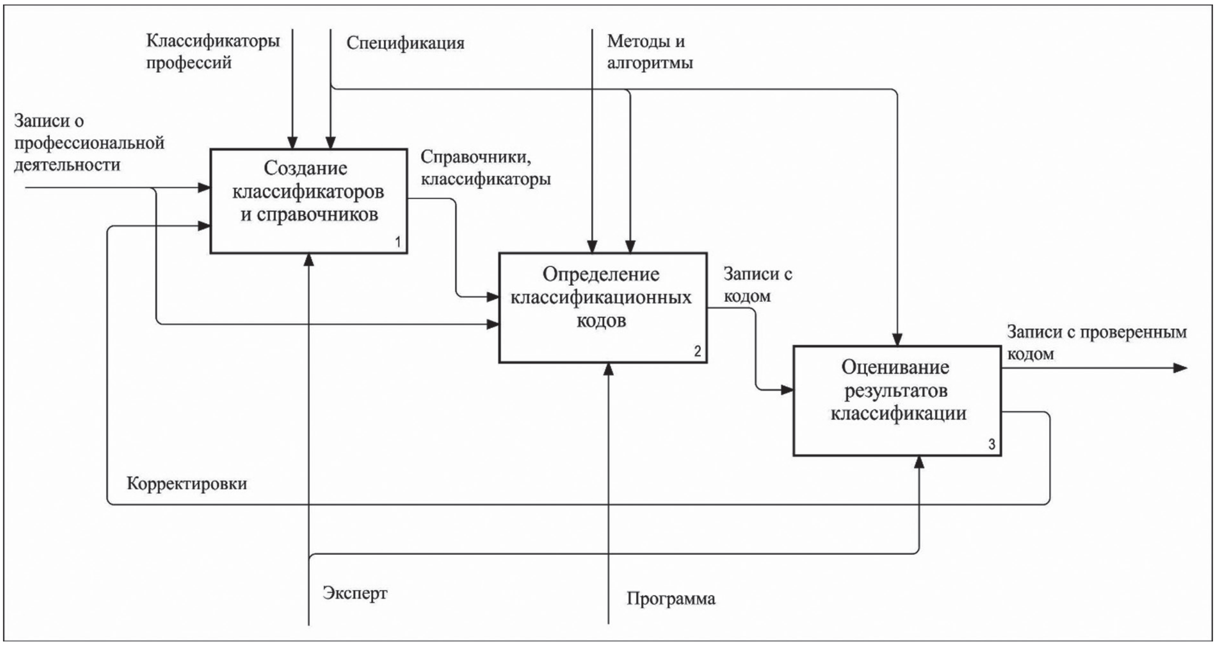

Рис. 1. Классификация записей о профессиональной деятельности

На рисунке 2 представлена диаграмма декомпозиции IDEF0 процесса определения классификационных кодов.

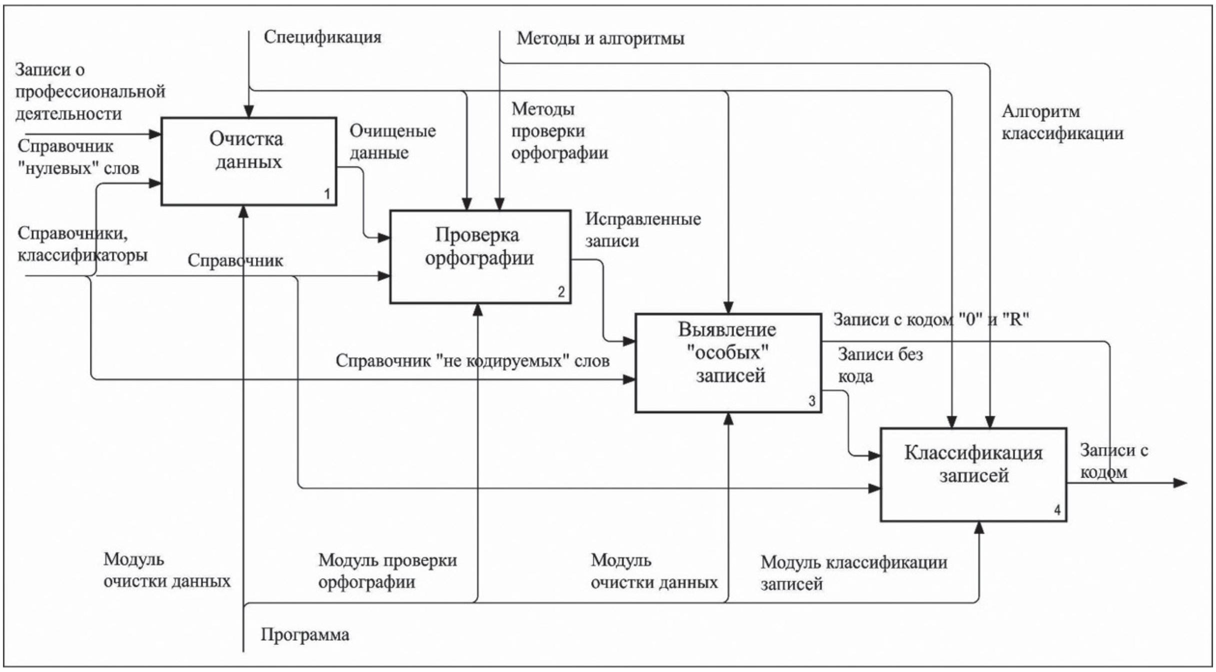

Рис. 2. Определение классификационных кодов для записей о профессиональной деятельности

Процесс определения классификационных кодов включает четыре этапа:

О2.1 — очистка данных;

О2.2 — проверка орфографии;

О2.3 — выявление «особых» записей;

О2.4 — классификация записей.

На этапе «Очистка данных» (*О2.1*) происходит удаление посторонних символов и цифр из описаний профессий, а также удаление «нулевых» слов с помощью специально созданного справочника. Как правило, «нулевые» слова присутствуют в качестве дополняющих или уточняющих характеристик описываемого вида деятельности. Например, описание «AFARMEROF 500 ACRES» после чистки преобразуется в «FARMERACRES», строка «APAINTEROFHOUSES» преобразуется в «PAINTERHOUSES», строка «ATAILOR, LEICESTER» преобразуется в «TAILOR».

На этапе «Проверка орфографии» (*О.2.2*) происходит выявление и устранение синтаксических ошибок в словах описания. Проверка ошибок выполняется на основе двух известных алгоритмов: алгоритме SPEEDCOP и алгоритме Левенштейна. В алгоритме SPEEDCOP используется ключ подобия, который строится путем объединения первой буквы слова, оставшихся уникальных согласных и уникальных гласных в порядке их встречаемости. Алгоритм основан наследующих положениях: (а) первая буква не является ошибкой; (б) согласные несут больше информации, чем гласные; (в) исходный порядок согласных, как правило, не нарушается; (г) удвоение и большинство перестановок букв не изменяет ключ. В алгоритме Левенштейна используется индекс «Расстояние Левенштейна» — минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую[5]. Оба алгоритма для выявления ошибок используют специальный словарь.

На этапе «Выявление «не кодируемых» записей» (*О2.3*) с помощью специально созданных справочников определяются записи двух категорий. К первой категории относятся записи, содержащие слова, указывающие на отсутствие занятости. Например, «UNEMPLOYED» (безработный), «BLACKSMITH RETIRED» (кузнец на пенсии), «FORMERLY A BLACKSMITH» (бывший кузнец). Ко второй категории относятся записи, содержащие слова, указывающие на родственные отношения (профессиональная деятельность описывается через третьих лиц). Например, «WIFE OF ABLACKSMITH» (жена кузнеца), «BLACKSMITH'S DAUGHTER» (дочь кузнеца). Выявление записей данного типа позволяет сохранить их для последующей обработки и при этом избежать ошибочной классификации и неверной интерпретации. В результате данного этапа записям первой категории присваивается код «0», записям второй категории — код «R».

На этапе «Классификация записей» (*О2.4*) реализуется два базовых алгоритма определения классификационных кодов, разработанных на основе теории принятия решений и байесовского подхода[6]. Для определения классификационных кодов используется справочник, подготовленный по результатам экспертной обработки объединенных данных, содержащий классификационные коды (c_k), слова, описывающие профессию (w_i) и их количество (n_{ik}). Описание алгоритмов представлено ниже.

1.1. Первый алгоритм определения классификационных кодов

Первый алгоритм заключается в определении классификационного кода для строки описания профессии на основе выявления наиболее значимого слова описания и его классификационного кода. Алгоритм включает следующие шаги:

Шаг 1. Извлечение слов из строки описания и формирование массива слов.

Строка = $\{w_1, w_2, \ldots, w_i, \ldots, w_n\}$.

Шаг 2. Расчет весового коэффициента слова

$$w_i = \sum_{k=1}^{K} \frac{n_{ik}}{N_k}, \qquad (1)$$

где w_i — весовой коэффициент i-ого слова;

n_{ik} — количество i-ого слова в k-ом коде;

N_k — общее количество слов в k-ом коде;

$k = \overline{1,K}$, K — количество кодов.

Шаг 3. Расчет коэффициента популярности слова (менее популярное слово важнее)

$$a_i = \frac{1}{N_i / N}, \qquad (2)$$

где a_i — коэффициент популярности i-ого слова;

N_i — общее количество i-ого слова;

N — общее количество слов.

Шаг 4. Расчет коэффициента порядка слова (первое слово в строке важнее)

$$\beta_i = \frac{1/(i+k)}{\sum_{i=1}^{n} 1/(i+k)}, \qquad (3)$$

где β_i — коэффициент порядка i-ого слова, $\sum_{i=1}^{n} \beta_i = 1$;

i — порядковый номер слова в строке, i = 1, 2, 3, .., n;

k — коэффициент отличия первого слова от остальных слов в строке, k = 1.

Шаг 5. Вычисление индекса слова.

Вариант 1:

$$Index_i^1 = w_i \cdot a_i \cdot \beta_i. \qquad (4)$$

Вариант 2:

$$Index_i^2 = w_i \cdot a_i, \qquad (5)$$

где $Index_i$ — индекс i-ого слова;

w_i — весовой коэффициент слова;

a_i — коэффициент популярности слова;

β_i — коэффициент порядка слова.

Шаг 6. Определение наиболее значимого слова в строке.

$$W_R = \arg\max(Index), \qquad (6)$$

где w_R — наиболее значимое слово строки — слово с максимальным значением Индекса.

Шаг 7. Определение классификационного кода для строки по значимому слову.

$$C^* = \arg\max(n_{w^*}), \qquad (7)$$

где n_{w^*} — количество значимого слова;

$C^*_{w_R}$ — классификационный код строки — код, где количество значимого слова максимально.

В результате работы первого алгоритма каждой записи (строке описания профессии) присваиваются два классификационных кода.

1.2. Второй алгоритм определения классификационных кодов

Второй алгоритм заключается в определении классификационного кода для строки описания профессии на основе комбинации классификационных кодов всех слов описания. Алгоритм включает следующие шаги:

Шаг 1. Извлечение слов из строки описания и формирование массива слов.

Строка = $\{w_1, w_2, \ldots, w_i, \ldots, w_n\}$.

Шаг 2. Расчет частоты слова для каждого кода.

$$v_{ik} = \frac{n_{ik}}{N_k}, \qquad (8)$$

где v_{ik} — частота i-ого слова в k-ом коде;

n_{ik} — количество i-ого слова в k-ом коде;

N_k — общее количество слов в k-ом коде.

Шаг 3. Расчет частоты слова с вероятностью ошибки.

$$v^*_{ik} = v_{ik} + P_0, \qquad (9)$$

где v^*_{ik} — частота i-ого слова в k-ом коде с вероятности ошибки P_0;

v_{ik} — частота i-ого слова в k-ом коде;

P_0 — вероятность ошибки, P_0 = 0.001.

Шаг 4. Расчет коэффициента популярности кода (профессии).

$$w_k = \frac{N_k}{N}, \qquad (10)$$

где w_k — коэффициент популярности кода;

N_k — количество слов в k-ом коде;

N — общее количество слов.

Шаг 5. Расчет весового коэффициента кода.

Вариант 1:

$$Q_k^1 = \prod_i v_{ik}^*. \qquad (11)$$

Вариант 2:

$$Q_k^2 = \prod_i v_{ik}^* \cdot (w_k), \qquad (12)$$

где Q_k — весовой коэффициент кода;

v^*_{ik} — частота i-ого слова в k-ом коде с вероятностью ошибки;

w_k — коэффициент популярности кода.

Шаг 6. Вычисление Индекса кода

Вариант 1:

$$P_k^1 = \frac{Q_k^1}{\sum_k Q_k^1} \qquad (13)$$

Вариант 2:

$$P_k^2 = \frac{Q_k^2}{\sum_k Q_k^2}, \qquad (14)$$

где P_k — Индекс кода;

Q_k — весовой коэффициент кода.

Шаг 7. Определение классификационного кода для строки описания.

$$Code^* = \arg\max(P), \qquad (15)$$

где $Code^*$ — классификационный код строки — код с максимальным значением Индекса.

В результате работы второго алгоритма каждой записи (строке описания профессии) присваиваются два классификационных кода.

1.3. Пример классификации записей

Рассмотрим применение алгоритмов определения классификационных кодов для следующей строки описания профессии: «COTTON RING SPINNER OVERLOOKER».

Согласно первому предложенному алгоритму на первом шаге выполняется извлечение слов из исходной строки описания. Формируется следующий набор слов:

w_1 = COTTON
w_2 = RING
w_3 = SPINNER
w_4 = OVERLOOKER

На втором шаге, в соответствии с формулой (1), рассчитывается весовой коэффициента каждого слова:

$$w_i = \frac{1173271}{2763615} + \frac{383881}{972759} + \frac{341070}{853174} + \frac{223748}{603840} + \frac{138974}{599958} + \ldots = 3.9029.$$

$$w_2 = \frac{25018}{972759} + \frac{81}{362432} + \frac{75}{91300} + \frac{54}{691820} + \frac{53}{362026} + \ldots = 0.0278.$$

$$w_3 = \frac{232811}{972759} + \frac{105110}{362026} + \frac{44325}{178115} + \frac{32965}{187026} + \frac{22976}{344594} + \ldots = 2.0625.$$

$$w_4 = \frac{12536}{603840} + \frac{9324}{2763615} + \frac{5697}{132254} + \frac{4409}{362026} + \frac{3951}{92807} + \ldots = 0.2341.$$

Здесь показан пример пяти слагаемых, соответствующих классификационным кодам, где слово встречается максимальное число раз: коды «551», «549», «550», «555» и «548» для слова «COTTON»; коды «549», «608», «382», «606» и «560» для слова «RING»; коды «549», «560», «559», «607» и «580» для слова «SPINNER» и коды «555», «551», «761», «560» и «572» для слова «OVERLOOKER».

На третьем шаге, в соответствии с формулой (2), рассчитывается коэффициент популярности слова:

$$a_1 = \frac{1}{2346263/150000514} = 63.9317;$$

$$a_2 = \frac{1}{25379/150000514} = 5910.4190;$$

$$a_3 = \frac{1}{495256/150000514} = 302.8747;$$

$$a_4 = \frac{1}{54756/150000514} = 2739.4350.$$

Согласно алгоритму наименее популярное слово имеет более высокое значение коэффициента. Таким образом, слово «RING» имеет наибольшее значение коэффициента, затем «OVERLOOKER» и «SPINNER», слово «COTTON» имеет наименьшее значение коэффициента.

На четвертом шаге, в соответствии с формулой (3), вычисляется коэффициент порядка. Слова строки описания имеют следующие значения коэффициента: $\beta_1 = 0.40$; $\beta_2 = 0.25$; $\beta_3 = 0.19$; $\beta_4 = 0.16$.

На пятом шаге, в соответствии с формулами (4) и (5), рассчитываются значения индекса слова.

Для варианта 1:
$Index^1_1 = 3{,}9029 \cdot 63{,}9317 \cdot 0.40 = 99{,}8076;$
$Index^1_2 = 0{,}0278 \cdot 5910{,}4190 \cdot 0.25 = 41{,}0774;$
$Index^1_3 = 2{,}0625 \cdot 302{,}8747 \cdot 0.19 = 118{,}6890;$
$Index^1_4 = 0{,}2341 \cdot 2739{,}4350 \cdot 0.16 = 102{,}6083.$
Для варианта 2:
$Index^2_1 = 3{,}9029 \cdot 63{,}9317 = 249{,}5190;$
$Index^2_2 = 0{,}0278 \cdot 5910{,}4190 = 164{,}3096;$
$Index^2_3 = 2{,}0625 \cdot 302{,}8747 = 624{,}6791;$
$Index^2_4 = 0{,}2341 \cdot 2739{,}4350 = 641{,}3017.$

На шестом шаге определяется наиболее значимое слово строки описания. Как видно из результатов предыдущего шага, для варианта 1 наиболее значимым словом является «SPINNER» со значением индекса 118,6890, для варианта 2 наиболее значимым словом является «OVERLOOKER» со значением индекса 641,3017. Следует заметить, что в варианте 2 разница между значениями индекса «OVERLOOKER» и «SPINNER» небольшая.

На седьмом шаге определяются классификационные коды для всей строки описания: по варианту 1 строке присваивается код «549», где значимое слово «SPINNER» представлено максимально (n_{549} = 232811), по варианту 2 строке присваивается код «555», где значимое слово «OVERLOOKER» представлено максимально (n_{555} = 12536).

Таким образом, в результате первого алгоритма рассматриваемой строке описания присваиваются два кода классификатора: «549» — «Прядение и производство хлопковых изделий» и «555» — «Хлопок и производство (неопределенно)».

Согласно второму предложенному алгоритму на первом шаге также выполняется извлечение слов из исходной строки описания. Формируется следующий набор слов:

w_1 = COTTON
w_2 = RING
w_3 = SPINNER
w_4 = OVERLOOKER

На втором шаге, в соответствии с формулой (8), для каждого слова описания рассчитывается частота по каждому существующему коду. Для наиболее значимых пяти кодов получены следующие результаты.

Слово «COTTON» максимально представлено в кодах «551», «549», «550», «555» и «548», следовательно, частота определяется так:

$$v_{1,551} = \frac{1173271}{2763615} = 0.4245; \; v_{1,549} = \frac{383881}{972759} = 0.3946;$$

$$v_{1,550} = \frac{341070}{853174} = 0.3998; \; v_{1,555} = \frac{223748}{603840} = 0.3705;$$

$$v_{1,548} = \frac{138974}{599958} = 0.2316.$$

Слово «RING» максимально представлено в кодах «549», «608», «382», «606» и «560», частота определяется как:

$$v_{2,549} = \frac{25018}{972759} = 0.0257; \; v_{2,608} = \frac{81}{36432} = 0.000223;$$

$$v_{2,382} = \frac{75}{91300} = 0.000821; \; v_{2,606} = \frac{54}{691820} = 0.000078;$$

$$v_{2,560} = \frac{53}{362026} = 0.00015.$$

Слово «SPINNER» максимально представлено в кодах «549», «560», «559», «607» и «580», частота определяется как:

$$v_{3,549} = \frac{232811}{972759} = 0.2393; \; v_{3,560} = \frac{105110}{362026} = 0.2903;$$

$$v_{3,559} = \frac{44325}{178115} = 0.2489; \; v_{3,607} = \frac{32965}{187026} = 0.1763;$$

$$v_{3,580} = \frac{22976}{344594} = 0.0667.$$

Слово «OVERLOOKER» максимально представлено в кодах «555», «551», «761», «560» и «572», частота определяется как:

$$v_{4,555} = \frac{12536}{603840} = 0.0208; \; v_{4,551} = \frac{9324}{2763615} = 0.0034;$$

$$v_{4,761} = \frac{5697}{132254} = 0.0431; \; v_{4,560} = \frac{4409}{362026} = 0.0122;$$

$$v_{3,572} = \frac{3951}{92807} = 0.0426.$$

Аналогично рассчитываются частоты по всем существующим кодам.

На третьем шаге алгоритма, в соответствии с формулой (9), рассчитывается частота слова с учетом вероятности ошибки. Для рассматриваемых кодов получены следующие значения:

$$v^*_{1,551} = 0.4255; \; v^*_{1,549} = 0.3956; \; v^*_{1,550} = 0.4008;$$
$$v^*_{1,555} = 0.3715; \; v^*_{1,548} = 0.2326.$$
$$v^*_{2,549} = 0.0267; \; v^*_{2,608} = 0.0012; \; v^*_{2,382} = 0.0018;$$
$$v^*_{2,606} = 0.0011; \; v^*_{2,560} = 0.0012.$$
$$v^*_{3,549} = 0.2403; \; v^*_{3,560} = 0.2913; \; v^*_{3,559} = 0.2499;$$
$$v^*_{3,607} = 0.1773; \; v^*_{3,580} = 0.0677.$$
$$v^*_{4,555} = 0.0218; \; v^*_{4,551} = 0.0044; \; v^*_{4,761} = 0.0441;$$
$$v^*_{4,560} = 0.0132; \; v^*_{4,572} = 0.0436.$$

В случае, когда код не содержит слово,

$$v^* = 0.001.$$

Аналогично рассчитывается частота по всем существующим кодам.

На четвертом шаге, в соответствии с формулой (10), рассчитывается коэффициент популярности кода. Для рассматриваемых кодов получены следующие значения:

$$w_{382} = 0.000609;\ w_{548} = 0.003999;\ w_{549} = 0.006485;$$
$$w_{550} = 0.005688;\ w_{551} = 0.018424;$$
$$w_{555} = 0.004026;\ w_{559} = 0.001187;\ w_{560} = 0.002420;$$
$$w_{572} = 0.000619;\ w_{580} = 0.002297;$$
$$w_{606} = 0.004612;\ w_{607} = 0.001247;\ w_{608} = 0.002416;$$
$$w_{761} = 0.000882.$$

Аналогично коэффициенты популярности рассчитываются для всех существующих классификационных кодов.

На пятом шаге, в соответствии с формулами (11) и (12), рассчитывается весовой коэффициент кода. Результаты для рассматриваемых кодов представлены в таблице 4.

Таблица 4

Пример расчета весового коэффициента кода

Code (k)	v^*_{1k}	v^*_{2k}	v^*_{3k}	v^*_{4k}	w_k	Q^1_k	Q^2_k
382	0.001	0.0018	0.001	0.001	0.000609	$1.8 \cdot 10^{-12}$	$1.1 \cdot 10^{-15}$
548	0.2326	0.001	0.001	0.001	0.003999	$2.3 \cdot 10^{-10}$	$9.3 \cdot 10^{-13}$
549	0.3956	0.0267	0.2403	0.001	0.006485	$2.2 \cdot 10^{-5}$	$1.4 \cdot 10^{-7}$
550	0.4008	0.001	0.001	0.001	0.005688	$4.0 \cdot 10^{-10}$	$2.2 \cdot 10^{-12}$
551	0.4255	0.001	0.001	0.0044	0.018424	$1.8 \cdot 10^{-9}$	$3.4 \cdot 10^{-11}$
555	0.3715	0.001	0.001	0.0218	0.004026	$6.3 \cdot 10^{-8}$	$2.5 \cdot 10^{-10}$
559	0.001	0.001	0.2499	0.001	0.001187	$2.5 \cdot 10^{-10}$	$2.9 \cdot 10^{-13}$
560	0.001	0.0012	0.2913	0.0132	0.002420	$4.6 \cdot 10^{-9}$	$1.1 \cdot 10^{-11}$
572	0.001	0.001	0.001	0.0436	0.000619	$1.6 \cdot 10^{-9}$	$9.8 \cdot 10^{-13}$
580	0.001	0.001	0.0677	0.001	0.002297	$5.5 \cdot 10^{-10}$	$1.3 \cdot 10^{-13}$
606	0.001	0.0011	0.001	0.001	0.004612	$1.1 \cdot 10^{-12}$	$4.9 \cdot 10^{-15}$
607	0.001	0.001	0.1773	0.001	0.001247	$1.8 \cdot 10^{-10}$	$2.2 \cdot 10^{-13}$
608	0.001	0.0012	0.001	0.001	0.002416	$1.2 \cdot 10^{-12}$	$2.9 \cdot 10^{-15}$
761	0.001	0.001	0.001	0.0441	0.000882	$4.4 \cdot 10^{-11}$	$3.9 \cdot 10^{-14}$

На шестом шаге, в соответствии с формулами (13) и (14), вычисляется индекс кода.

Для варианта 1:

$$P^1_{382} = 6.2 . 10^{-8};\ P^1_{548} = 7.8 . 10^{-6};\ P^1_{549} = 0.771213;$$
$$P^1_{550} = 1.4 . 10^{-5};\ P^1_{551} = 6.3 . 10^{-5};$$
$$P^1_{555} = 0.002121;\ P^1_{559} = 8.4 . 10^{-6};\ P^1_{560} = 0.000155;$$
$$P^1_{572} = 5.4 . 10^{-5};\ P^1_{580} = 1.9 . 10^{-5};$$
$$P^1_{606} = 3.6 . 10^{-8};\ P^1_{607} = 5.9 . 10^{-6};\ P^1_{608} = 4.1 . 10^{-8};$$
$$P^1_{761} = 1.5 . 10^{-6}.$$

Для варианта 2:

$$P^2_{382} = 7.5 . 10^{-9};\ P^2_{548} = 6.3 . 10^{-6};\ P^2_{549} = 0.997784;$$
$$P^2_{550} = 1.5 . 10^{-5};\ P^2_{551} = 0.000231;$$
$$P^2_{555} = 0.001703;\ P^2_{559} = 2.0 . 10^{-6};\ P^2_{560} = 7.5 . 10^{-5};$$
$$P^2_{572} = 6.6 . 10^{-6};\ P^2_{580} = 8.6 . 10^{-6};$$
$$P^2_{606} = 3.4 . 10^{-8};\ P^2_{607} = 1.5 . 10^{-6};\ P^2_{608} = 1.9 . 10^{-8};$$
$$P^2_{761} = 2.6 . 10^{-7}.$$

Следует заметить, что нормализация выполняется на основе весовых коэффициентов всех существующих классификационных кодов. Аналогично рассчитывается индекс для всех существующих кодов.

На седьмом шаге определяются классификационные коды для строки описания. Принимая во внимание результаты расчетов по всем существующим кодам, определяются следующие пять кодов с максимальными значениями индекса.

Для варианта 1:

‘549’ – $P^1_{549} = 0{,}771213$; ‘555’ – $P^1_{555} = 0{,}002121$;
‘520’ – $P^1_{520} = 0{,}000273$;
‘560’ – $P^1_{560} = 0{,}000155$; ‘525’ – $P^1_{525} = 0{,}000149$;
‘571’ – $P^1_{571} = 0{,}000098$.

Для варианта 2:

‘549’ – $P^2_{549} = 0{,}997784$; ‘555’ – $P^2_{555} = 0{,}001703$;
‘551’ – $P^2_{551} = 0{,}000231$;
‘560’ – $P^2_{560} = 0{,}000075$; ‘550’ – $P^2_{550} = 0.000015$;
‘571’ – $P^2_{571} = 0.000012$.

Как видно из результатов, значение для кода «549» существенно больше значений остальных кодов в двух вариантах. По результатам второго алгоритма рассматриваемой строке присваивается код «549» — «Прядение и производство хлопковых изделий» для двух вариантов.

Таким образом, в результате применения двух алгоритмов каждой записи было присвоено четыре классификационных кода. Далее перед экспертами встает другая задача — понять, какой из найденных четырех кодов «наиболее правильный». Оценка результатов классификации показала, что 1% записей имеют четыре одинаковых кода; 20% записей имеют одинаковые коды в трех случаях из четырех; 63% записей имеют одинаковые коды для двух вариантов первого алгоритма; 5,5% записей имеют одинаковые коды для двух вариантов второго алгоритма; 4% записей имеют два одинаковых кода, полученных двумя алгоритмами в разных сочетаниях и 5,5% записей имеют абсолютно разные коды во всех четырех случаях.

2. МЕТОД СТАНДАРТИЗАЦИИ ЗАПИСЕЙ О МЕСТЕ РОЖДЕНИЯ

Рассматривая задачу стандартизации записей о месте рождения, следует отметить, что информация в книгах переписи представляет собой данные трех уровней, соответствующих территориально-административной структуре Великобритании: *country* (страна), *county* (графство) и *parish* (приход). Поэтому стандартизация записей выполняется в соответствии с этой географической иерархией, хотя порядок данных в записях не всегда строго соответствует этим уровням (приход может быть записан как графство, а графство как страна, и наоборот).

Учитывая структуру записей и опираясь на уже полученные ранее результаты по расшифровке данных переписей 1881 и 1911 гг.[7], предварительно, анализируя данные о стране и графстве, для каждой записи были установлены коды графств. Эти коды позволяют идентифицировать записи, относящиеся к территориям Англии, Уэльса и Шотландии, для последующей обработки данных на уровне приходов.

Также следует отметить некоторые особенности, определяющие стратегию стандартизации. Приходы, охватывающие обширные пощади, могут включать в себя более мелкие населенные пункты (города, деревни, поселения) или представлять собой агломерацию других приходов. Например, приход *Hatfield Board Oak* в графстве *Essex* включает в себя два поселения *Bush End* и *Hatfield Heath*, а город *Southend-on-Sea* представляет собой агломерацию четырех приходов: *Prittlewell*, *Leigh*, *Southchurch* и *Eastwood*. Кроме того, возможно было изменение названий некоторых населенных пунктов, например, *Hatfield Board Oak* был также известен как *Hatfield Regis*, названный в честь королевского леса, занимающего большую часть территории. Таким образом, разрабатываемый метод стандартизации должен давать возможность не только определять «правильное» название населенных пунктов, но и обеспечивать их привязку к «родительским» приходам.

Метод стандартизации записей о месте рождения основан на подходе, аналогичном классификации записей о профессиональной деятельности. На основе SADT-методологии разработана функциональная модель процессов. На рисунке 3 представлена контекстная диаграмма IDEF0 процесса стандартизации.

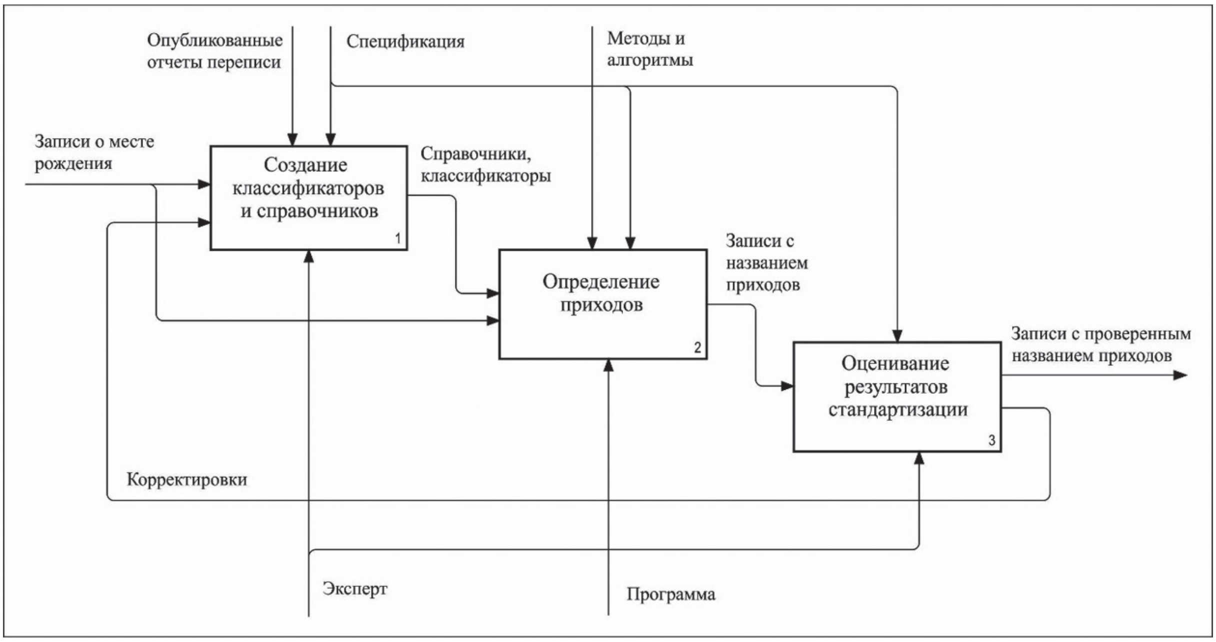

Рис. 3. Стандартизация записей о месте рождения

Процесс стандартизации записей включает три основных этапа:

B1 — создание классификаторов и справочников;

B2 — идентификация приходов;

B3 — оценивание результатов стандартизации.

На первом этапе (*B1*) осуществляется разработка и модификация необходимых справочников и классификаторов. Данный этап выполняется экспертами на основе существующих географических справочников и опубликованных отчетов.

Для данной задачи стандартизации записей разработан справочник территорий, содержащий коды графств (CNTI); названия приходов (STD_PAR); названия более мелких населенных пунктов, входящих в состав приходов (PLACE), и весовые коэффициенты населенных пунктов, учитывающие численность населения и используемые при разрешении конфликтов в процессе стандартизации (WEIGHT). Также разработаны дополнительные справочники, используемые на этапах идентификации приходов.

На втором этапе (*B2*) для исходной строки с описанием населенного пункта определяется название прихода. Данный этап реализуется программой на основе разработанных методов и алгоритмов с использованием созданных справочников и классификаторов.

На третьем этапе (*B3*) выполняется оценивание результатов стандартизации с внесением необходимых корректировок в справочники и алгоритмы, формирование итоговых результатов.

На рисунке 4 представлена диаграмма декомпозиции IDEF0 процесса идентификации прихода.

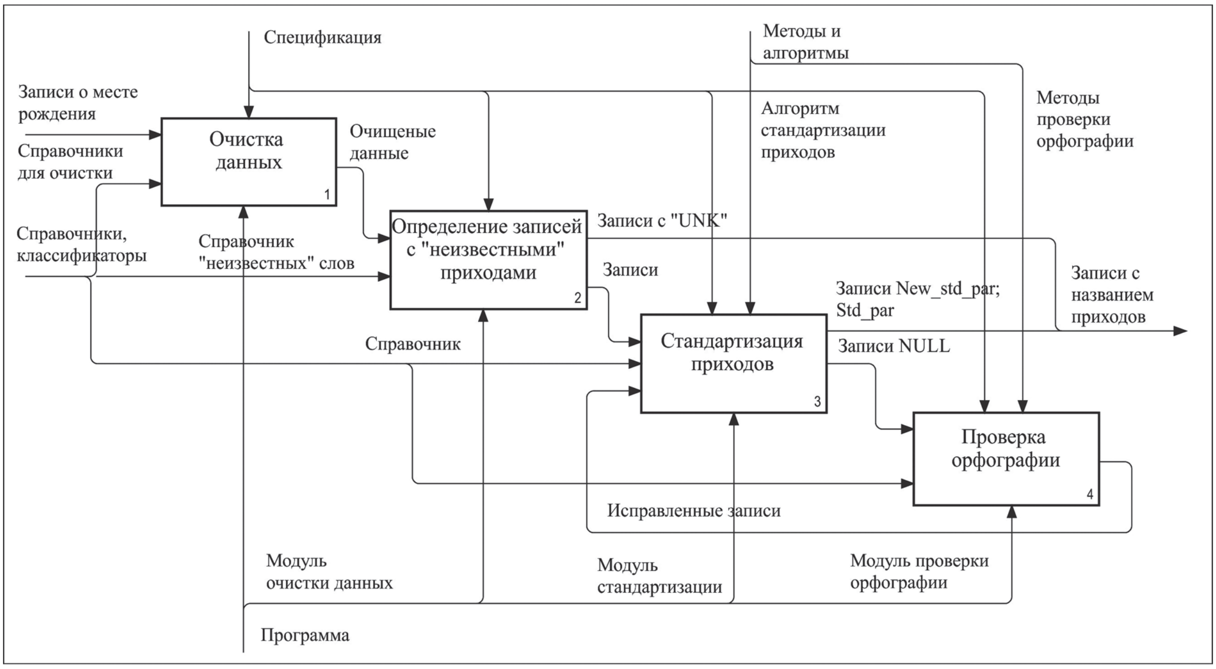

Рис. 4. Идентификация приходов для записей о месте рождения

Процесс идентификации приходов для исходной строки с описанием населенного пункта включает четыре этапа:

B2.1 — очистка данных;

B2.2 — определение записей с «неизвестными» приходами;

*B2.*3 — стандартизация приходов;

B2.4 — проверка орфографии.

На этапе «Очистка данных» (*B2.1*) происходит предварительная обработка исходных данных в строке описания места рождения: 1) преобразование аббревиатур и сокращений, например, GT преобразуется в GREAT, ST — в SAINT, LT — в LITTLE; 2) удаление посторонних символов, цифр, одиночных букв, удаление наименований стран (например, ENGLAND, BRITISH, SCOTLAND), удаление наименований графств (например, ARGYLLSHIRE, BEDFORDSHIRE, CAMBRIDGESHIRE), а также удаление слов, не содержащих информацию о названии населенного пункта (например, PLACE, FROM, RESIDENT, NEAR); 3) удаление элементов строки, относящихся к описанию адреса, с помощью слов-указателей, таких как: ROAD, STREET, LANE, RD, ST. Очистка данных выполняется на основе специально созданных справочников.

На этапе «Определение записей с «неизвестными» приходами» (*B2.2*) происходит обнаружение записей, не содержащих названий населенных пунктов в связи с тем, что многие люди могли не знать точно место своего рождения. Выявление записей с «неизвестными» приходами выполняется с помощью специально подготовленного справочника, содержащего ключевые слова-указатели, такие как: UNKNOWN, NOT KNOWN, NOT NAME, NK, BLANK. В результате данного этапа обнаруженным записям присваивается значение «UNK» («неизвестный»).

На этапе «Стандартизация приходов» (*B2.3*) реализуется алгоритм, основанный на сопоставлении исходного описания (строки с названием населенного пункта) со справочными данными. Метод предполагает три типа результата: 1) для исходной записи определяется стандартное название прихода из справочника («STD_PAR»); 2) для исходной записи определяется стандартное название прихода, но меняется код графства («NEW_STD_PAR»); 3) для исходной записи стандартное название прихода не определено («NULL»). Описание алгоритма представлено ниже.

На этапе «Проверка орфографии» (*B2.4*) для записей типа «NULL» выполняется проверка и исправление синтаксических ошибок. На данном этапе, как и в случае классификации профессий, проверка орфографии выполняется на основе двух алгоритмов: алгоритме SPEEDCOP и алгоритме Левенштейна. При этом для исправления исходных названий используются населенные пункты, расположенные в том же графстве. После исправления записи возвращаются на этап *B2.3* для повторной обработки. В результате второй итерации на этапе *B2.3* формируются два типа результатов: «STD_PAR» и «NEW_STD_PAR».

2.1. Алгоритм стандартизации приходов для записей о месте рождения

На рисунках 5 и 6 представлено описание механизма идентификации в виде диаграммы событий управляющей модели ARIS (ARchitecture of Integrated Information Systems)[8]. Диаграмма описывает процесс в виде последовательности событий, действий и других вспомогательных объектов (в данном случае, справочников). События (на диаграмме изображаются в виде шестиугольников) описывают обстоятельства или условия, при которых выполняются действия. Действия (на диаграмме изображаются в виде прямоугольников) описывают преобразования из начального состояния в конечное состояние. Правила логики («И», «ИЛИ», «исключающее ИЛИ») используются для логической связки между событиями и действиями.

Согласно предложенному алгоритму, идентификация выполняется в два этапа: сначала реализуется проверка целой строки описания места рождения, затем реализуется проверка отдельных слов, извлеченных из строки описания.

Первый этап алгоритма включает следующие шаги (рис. 5):

Шаг 1.1. Проверка уникального названия прихода. Необходимость такой проверки вызвана тем, что многие приходы имеют одинаковые названия. Если исходная строка совпадает с названием уникального прихода в справочнике (STD_PAR) и совпадает код графства (CNTI), то строке присваивается название прихода (STD_PAR).

Шаг 1.2. Проверка уникального названия населенного пункта. Если исходная строка совпадает с названием уникального населенного пункта (PLACE) и совпадает код графства (CNTI), то строке присваивается название прихода, связанного с найденным населенным пунктом (STD_PAR).

Шаг 1.3. Создание «сжатой» строки (удаление пробелов).

Шаг 1.4. Проверка уникального названия прихода в «сжатой» строке. Если «сжатая» строка совпадает с названием уникального прихода в справочнике (STD_PAR) и совпадает код графства (CNTI), то строке присваивается название прихода (STD_PAR).

Шаг 1.5. Проверка уникального названия населенного пункта в «сжатой» строке. Если «сжатая» строка совпадает с названием уникального населенного пункта в справочнике (PLACE) и совпадает код графства (CNTI), то строке присваивается название прихода, связанного с найденным населенным пунктом (STD_PAR).

Шаг 1.6. Проверка населенных пунктов, названия которых совпадают с названиями графств. Проверка выполняется с помощью справочника, содержащего названия населенных пунктов, эквивалентные названию графств, например, BUCKINGHAM, BEDFORD, CAMBRIDGE, DERBY, LEICESTER, OXFORD. Если строка содержит такое название, то оно удаляется, и проверка строки начинается сначала. Иначе первый этап заканчивается.

Второй этап алгоритма включает следующие шаги (рис. 6):

Шаг 2.1. Извлечение слов из строки и формирование массива слов.

Шаг 2.2. Проверка уникального названия прихода. Если исходное слово совпадает с названием уникального прихода в справочнике (STD_PAR) и совпадает код графства (CNTI), то присваивается название прихода (STD_PAR).

Шаг 2.3. *Идентификация прихода.* Идентификация основана на сравнении исходного слова с названием населенного пункта в справочнике (PLACE). На данном шаге реализуется несколько условий:

— Если исходное слово совпадает с уникальным названием населенного пункта в справочнике (PLACE) и совпадает код графства (CNTI), то присваивается название прихода, связанного с найденным населенным пунктом (STD_PAR).

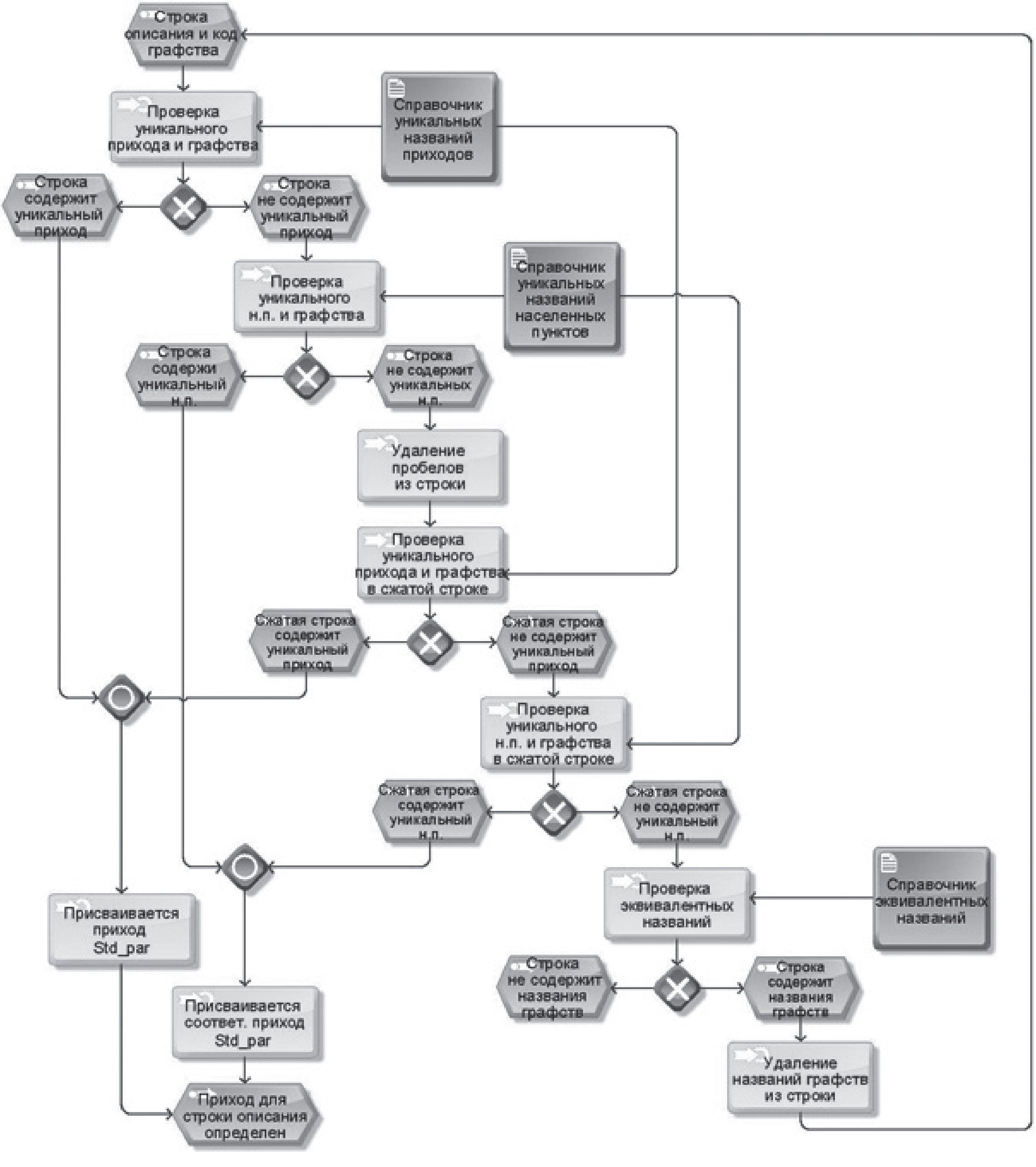

Рис. 5. Идентификация приходов: проверка строки описания

— Если исходное слово совпадает с названием населенного пункта в справочнике (PLACE) и совпадает код графства (CNTI), но им соответствует несколько названий приходов, то присваивается название прихода (STD_PAR), который имеет наибольшее значение весового коэффициента (WEIGH).

— Если исходное слово совпадает с уникальным названием населенного пункта в справочнике (PLACE), но не совпадает код графства (CNTI), то устанавливается новое название графства (NEW_CNTI) и присваивается соответствующее название прихода (NEW_ STD_PAR).

— Если исходное слово совпадает с названием населенного пункта в справочнике (PLACE), которому соответствует несколько графств (CNTI), то выбирается графство, которое ближе всех расположено к исходному (по таблице расстояний), уста-

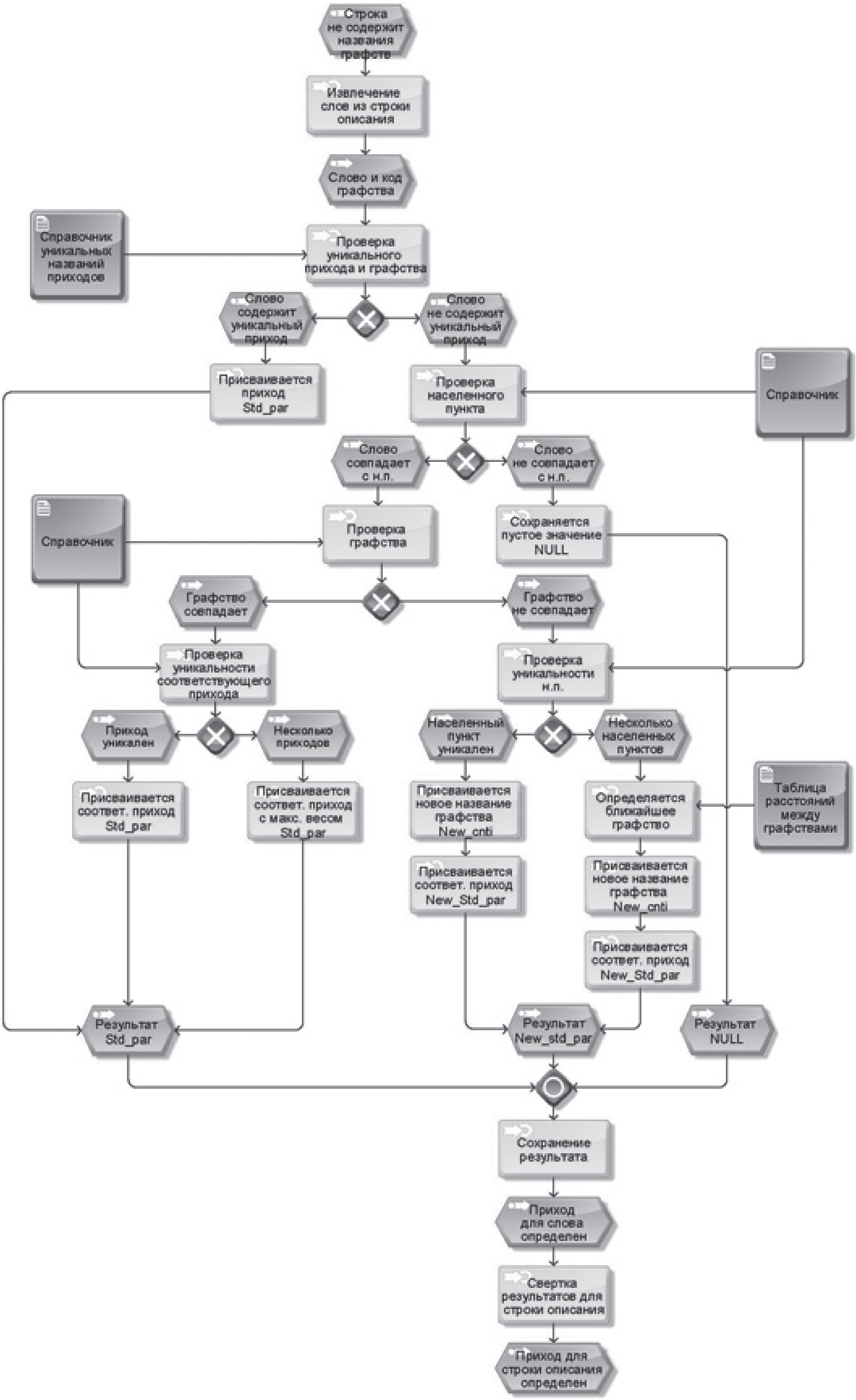

Рис. 6. Идентификация приходов: проверка слов из строки описания

навливается новое название графства (NEW_CNTI) и присваивается соответствующее название прихода (NEW_ STD_PAR).

— Если исходное слово не совпадает ни с одним названием населенного пункта в справочнике, то сохраняется пустое значение (NULL).

Шаг 2.4. Свертка результатов и определение прихода для исходной строки. По результатам, полученным для отдельных слов, определяется итоговое название прихода для исходной строки описания по следующим правилам:

— Если ответы соответствуют разным типам, то результат формируется по приоритету: STD_PAR; NEW_ STD_PAR; NULL.

— Если существует два и более ответа одного типа (STD_PAR или NEW_ STD_PAR), то в качестве итогового выбирается название прихода с минимальным значением весового коэффициента.

Строки, получившие тип NULL, подлежат проверке и исправлению орфографических ошибок и повторной обработке. Строки, получившие ответ типа NEW_ STD_PAR, подлежат экспертному анализу. Для таких строк характерны два вида ошибок: 1) в названии исходного графства, 2) в названии населенного пункта. В первом случае эксперт принимает новое название графства (NEW_CNTI) и соответствующее ему новое название прихода (NEW_STD_PAR), во втором случае исправляется название населенного пункта, сохраняется название графства (CNTI) и определяется название соответствующего прихода (STD_ PAR). В результате стандартизации для каждой записи о месте рождения определено стандартное название прихода.

2.2. Пример стандартизации приходов

Рассмотрим применение предложенного алгоритма идентификации приходов для следующих записей:

1. HURDSFIELD | CHS
2. GREAT HASELEY | OXF
3. CAMBRIDGE RAMPTON | CAM
4. LINCOLN TATTESHALL | LIN
5. MIDDLETON | SAL
6. CLIFTON | SOM
7. LLANFAIRFERNDALE | GLA

Каждая запись содержит строку описания места рождения и код графства.

Шаг 1.1. Строка 2 содержит уникальное название прихода, следовательно, присваивается название прихода GREATHASELEY.

Шаг 1.2. Строка 1 содержит уникальное название населенного пункта, следовательно, присваивается название прихода MACCLESFIELD (населенный пункт HURDSFIELD входит в состав прихода MACCLESFIELD).

После шагов 1.1 и 1.2 получены следующие результаты:

1. HURDSFIELD | CHS — MACCLESFIELD
2. GREAT HASELEY | OXF — GREAT HASELEY
3. CAMBRIDGE RAMPTON | CAM
4. LINCOLN TATTESHALL | LIN
5. MIDDLETON | SAL
6. CLIFTON | SOM
7. LLANFAIR FERNDALE | GLA

Шаг 1.3. Формируются следующие «сжатые» строки:

3. CAMBRIDGERAMPTON | CAM
4. LINCOLNTATTESHALL | LIN
5. MIDDLETON | SAL
6. CLIFTON | SOM
7. LLANFAIRFERNDALE | GLA

Шаг 1.4. Совпадений с уникальными названиями приходов в «сжатых» строках не найдено.

Шаг 1.5. Совпадений с уникальными названиями населенных пунктов в «сжатых» строках не найдено.

Шаг 1.6. Строки 3 и 4 содержат названия населенных пунктов, эквивалентные названиям графств: CAMBRIDGE в строке 3 и LINCOLN в строке 4. После удаления эквивалентных названий получены следующие результаты:

3. RAMPTON | CAM
4. TATTESHALL | LIN
5. MIDDLETON | SAL
6. CLIFTON | SOM
7. LLANFAIRFERNDALE | GLA

Шаг 1.1 (второй цикл). Строка 3 содержит уникальное название прихода, следовательно, присваивается название прихода RAMPTON.

После завершения первого этапа алгоритма получены следующие результаты:

1. HURDSFIELD | CHS — MACCLESFIELD | CHS
2. GREAT HASELEY | OXF — GREAT HASELEY | OXF
3. RAMPTON | CAM — RAMPTON| CAM
4. TATTESHALL | LIN
5. MIDDLETON | SAL
6. CLIFTON | SOM
7. LLANFAIR FERNDALE | GLA

Строки 4, 5, 6 и 7 переходят на второй этап алгоритма — проверку отдельных слов из строки описания.

Шаг 2.1. Получены следующие слова:

4. TATTERSHALL | LIN
5. MIDDLETON | SAL
6. CLIFTON | SOM

7a. LLANFAIR | GLA
7b. FERNDALE | GLA

Шаг 2.2. Строка 7a содержит уникальное название прихода, следовательно, присваивается название прихода LLANFAIR.

Шаг 2.3.

— Строка 7б содержит уникальное название населенного пункта, следовательно, присваивается название прихода RHONDDA (населенный пункт FERNDALE входит в состав прихода RHONDDA).

— Сочетание MIDDLETON | SAL в строке 5 не уникальное, населенный пункт MIDDLETON существует в двух приходах графства SHROPSHIRE (SAL): BITTERLEY (вес=168) and OSWESTRY (вес=1307). Следовательно, строке 5 присваивается название прихода OSWESTRY, имеющего наибольший вес.

— Строка 6 содержит название населенного пункта CLIFTON, которое не существует в исходном графстве SOMERSET (SOM). В справочнике семь графств с населенным пунктом CLIFTON (WES, WOR, BDF, DBY, DEV, GLS, LAN), ближайшим из которых к исходному графству является GLOUCESTER (GLS). Следовательно, устанавливается новое название графства — GLS и новое название прихода — CLIFTON.

— Строка 4 содержит название, которое не соответствует ни одному населенному пункту в справочнике. Следовательно, строке присваивается пустое значение.

В результате шага 2.3. получены следующие результаты:

4. TATTERSHALL | LIN —
5. MIDDLETON | SAL– OSWESTRY | SAL
6. CLIFTON | SOM–CLIFTON | GLS
7a. LLANFAIR | GLA — LLANFAIR | GLA (вес 16)
7b. FERNDALE | GLA — RHONDDA | GLA (вес 1307)

Шаг 2.4 Строка 7 имеет два варианта ответа: LLANFAIR | GLA (вес =16) и RHONDDA | GLA (вес =1307). Следовательно, строке присваивается название прихода LLANFAIR, имеющего наименьшее значение весового коэффициента.

После завершения первого и второго этапов алгоритма идентификации получены следующие результаты:

1. HURDSFIELD | CHS — MACCLESFIELD | CHS
2. GREAT HASELEY | OXF — GREAT HASELEY | OXF
3. CAMBRIDGE RAMPTON | CAM — RAMPTON | CAM
4. LINCOLN TATTESHALL | LIN — (null)
5. MIDDLETON | SAL — OSWESTRY | SAL
6. CLIFTON | SOM–CLIFTON | GLS (новый)
7. LLANFAIRFERNDALE | GLA — LLANFAIR | GLA

Строка 4, имеющая результат типа «NULL», подлежит проверке орфографии и повторной обработке по всем шагам алгоритма. После чего строка 4 получает результат: TATTERSHALL | LIN — ATTERSHALL| LIN. Строка 6, имеющая результат типа «New_std_par», подлежит экспертному анализу. После чего строка 6 получает результат: CLIFTON | SOM–CLIFTON | GLS.

ЗАКЛЮЧЕНИЕ

Задачи стандартизации и классификации текстуальных записей о месте рождения и профессиональной деятельности, полученных по данным переписи, актуальные и востребованы. Как правило, данные проблемы решаются путем обработки данных вручную или полуавтоматически, используя средства поиска, сортировки или фильтрации, позволяющие оптимизировать процессы сравнения и кодирования записей. Значительный объем и особый характер исходных данных в данной работе потребовали иного подхода, основанного на автоматической обработке. С учетом сложной структуры и многовариантности описаний на основе теории принятия решений и технологии структурного анализа, разработаны алгоритмы формирования классификационных кодов профессий и идентификации географических районов. Используя экспертные знания, анализируя вручную небольшой, но репрезентативный набор данных, разработанные алгоритмы были применены для автоматической обработки миллионов записей. При этом алгоритмы сравнения использовались в комбинации с алгоритмами преобразования, проверки синтаксических ошибок и расчета коэффициентов: коэффициента популярности профессий — для классификации видов деятельности, коэффициентов близости и коэффициентов значимости территорий — для стандартизации мест рождения. Такой сложный набор шагов позволил обработать большой объем данных. Несмотря на погрешности, связанные с непростым характером данных, избежать которые можно только владея глубокими знаниями о географических особенностях территорий или специальными знаниями о профессиональной деятельности в различных районах страны, полученная база данных вместе с разработанной методикой представляют важнейший результат и могут успешно применяться для решения аналогичных задач стандартизации.

ПРИМЕЧАНИЯ

[1] Lawton R. The Census and Social Structure. An Interpretative Guide to Nineteenth Century Censuses for England and Wales. London, 1978; Wrigley E. A. Nineteenth-century society. Essays in the use of quantitative methods for the study of social data. Cambridge, 1972.; Mills D. R., Schürer K. Local Communities in the Victorian Census Enumerators' Books. Oxford, 1996; Higgs E. Making Sense of the Census Revisited. Census Records for England and Wales, 1801–1901 — a Handbook for Historical Researchers, London: The National Archives and Institute of Historical Research. London, 2005.

[2] Project "Mining Microdata: economic opportunity and spatial mobility in Britain, Canada and The United States, 1850–1911" [Electronic Resource]. URL: http://www.miningmicrodata.org; Schürer K., Higgs E. Integrated Census Microdata (I–CeM): 1851–1911 [Electronic Resource]. URL: http://www.essex.ac.uk/history/research/icem; Colchester, Essex: UK Data Archive [distributor], April 2014, SN: 7481 [Electronic Resource]. URL: http://dx.doi.org/10.5255/UKDA-SN-7481–1; Higgs E., Jones C., Schürer K., Wilkinson A. The Integrated Census Microdata (I–CeM) Guide, Colchester, 2013 [Electronic Resource]. URL: http://www.essex.ac.uk/history/researchicem/documents/icem_guide.pdf).

[3] Marca D. A., McGowan C. L. SADT: Structured Analysis and Design Technique, McGraw-Hill, New York, 1987; Davis W. S. Business systems analysis and design // Business & Economics, 1994.

[4] I–CeM Occupational Matrix [Electronic Resource]. URL: http, www.essex.ac.uk/history/research/i cem/documentation.html; Schurer K. Woollard M. 1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version). Colchester, Essex: UK Data Archive [distributor], November 2000. SN: 4177 [Electronic Resource]. URL: http://dx.doi.org/10.5255/UKDA-SN-4177–1; Woollard M. The classification of occupations in the 1881 census of England and Wales, Colchester, 199); Eames C. Eames R. A Computer Perspective. Background to the Computer Age, London, 199); Austrian G. Hollerith H. Forgotten Giant of Information Processing, New York, 1982; Anderson M. J. The American Census. A Social History, London, 1988; BPP 1911, CVII, General Report with Appendices, Appendix B; Higgs E. The statistical Big Bang of 1911: ideology, technological innovation and the production of medical statistics, Social History of Medicine. 1996. N 9. P. 409–26; Higgs E. The Information State in England: the Central Collection of Information on Citizens, 1500–2000. London, 2004; Pollock J., Zamora A. Spelling correction in scientific and scholarly, Communications of the ACM. 1984. V. 27. N.4. P. 358–368.

[5] Schierle M., Schulz S., Ackermann M. From spelling correction to text cleaning — using context information, Data Analysis, Machine Learning and Applications, series Studies in Classification, Data Analysis, and Knowledge Organization. 2008. V.27. N.4. pp. 397–404.

[6] Berger J. O. Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics. Springer-Verlag, 1985. 617 p.

[7] Schurer, K., "The 1891 census and local population studies", Local Population Studies. 1991. N 47. P. 16–29; Woollard, M. The classification of occupations in the 1881 census of England and Wales, Colchester, 1999.

[8] Whitten J., Barlow V., Bentley L. Systems Analysis and Design Methods. McGraw-Hill Professional. 1997. 896 p.; Hommes B.-J. The Evaluation of Business Process Modeling Techniques. TU Delft. 2004. 137 p.