VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INTELIGENTNÍCH SYSTÉMŮ
DEPARTMENT OF INTELLIGENT SYSTEMS

SPOJOVÁNÍ ZÁZNAMŮ V GENEALOGICKÝCH DATECH

RECORD LINKAGE IN GENEALOGICAL DATA

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **JAN ŠORM**
AUTHOR

VEDOUCÍ PRÁCE **doc. Ing. FRANTIŠEK ZBOŘIL, Ph.D.**
SUPERVISOR

BRNO 2019

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav inteligentních systémů (UITS) Akademický rok 2018/2019

Zadání bakalářské práce

Student: **Šorm Jan**
Program: Informační technologie
Název: **Spojování záznamů v genealogických datech**
Record Linkage in Genealogical Data
Kategorie: Umělá inteligence

Zadání:

1. Prostudujte metody slučování dat v historických dokumentech a seznamte se s problematikou vytváření genealogických modelů
2. Pro dodanou sadu přepsaných matričních záznamů v elektronické podobě identifikujte problémy při identifikaci rolí, které v těchto záznamech mohou osoby sehrávat a seznamte se se způsoby zápisů jmen, stavu, povolání a bydlišť osob v 17. až 19. století.
3. Zvolte a otestujte metody, které by umožnily identifikovat osobu, která je zapsána různými způsoby, včetně toho, kdy některé části jsou zapsány v různých jazycích.
4. Ověřte funkčnost Vámi navržených a implementovaných metod při použitích v aplikacích pro vytváření genealogických modelů.
5. Zhodnoťte dosažené výsledky, diskutujte další možný postup a vytvořte plakát formátu A2, který bude výsledky práce demonstrovat.

Literatura:

- 1. Fellegi, I.,P., Sunter, A.,B.: A theory for record linkage, 1969

Pro udělení zápočtu za první semestr je požadováno:

- První dva body zadání

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Zbořil František, doc. Ing., Ph.D.**
Vedoucí ústavu: Hanáček Petr, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 15. května 2019
Datum schválení: 1. listopadu 2018

Abstrakt

Hlavním cílem této bakalářské práce je studium genealogických dat, zjištění možných problému při jejich slučování a implementace metod pro toto slučování dat. V této práci se bude především studovat problematika třídění podobných tvarů jmen do společných tříd. Tento problém se bude studovat zejména proto, že v každém matričním záznamu hrají nejdůležitější roli jména a příjmení dotčených osob a jejich příbuzných. V práci tedy bude rozebráno několik metrik pro výpočet vzdálenosti mezi dvěma řetězci. Dále pak pro tyto metriky bude provedeno několik experimentů, které budou mít za cíl roztřídit jména do tříd s co nejmenším počtem chyb. Na základě těchto výsledků pak budou provedeny i experimenty pro samotné slučování jednotlivých genealogických záznamů.

Abstract

The main aim of this thesis is to study genealogical data, to find out possible problems in their merging and to implement methods for this data merging. In this thesis, it will be studied the problem of classifying similar names into common classes. This problem will be studied mainly because people's names and surnames play the most important role in every registry entry. In this thesis, it will be analyzed several metrics for calculating the distance between two strings. In addition, several experiments will be done for these metrics to classify names into classes with as few errors as possible. Based on these results, experiments for record linkage will be performed.

Klíčová slova

genealogie, matrika, záznamy, slučování, řetězce, vzdálenosti, třídy, C++

Keywords

genealogy, register, records, merging, strings, distances, classes, C++

Citace

ŠORM, Jan. *Spojování záznamů v genealogických datech.* Brno, 2019. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. Ing. František Zbořil, Ph.D.

Spojování záznamů v genealogických datech

Prohlášení

Prohlašuji, že jsem tuto semestrální práci vypracoval samostatně pod vedením pana doc. Ing. Františka Zbořila, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jan Šorm
14. května 2019

Poděkování

Děkuji doc. Ing. Františku Zbořilovi, Ph.D. za vedení mé bakalářské práce a věcné připomínky při konzultacích.

Obsah

1 Úvod **3**

2 Teorie **4**
2.1 Teorie Fellegi-Sunter 4
2.2 Srovnání záznamů dle data záznamu 6
2.2.1 Rozdělení pravděpodobnosti spojité náhodné veličiny 6
2.2.2 Normální rozdělení 6
2.2.3 Historické souvislosti 7
2.3 Metody pro porovnávání řetězců 8
2.3.1 Hammingova vzdálenost 8
2.3.2 Levenshteinova vzdálenost 8
2.3.3 Jarova vzdálenost 9
2.3.4 Jaro-Winklerova vzdálenost 10
2.3.5 Soundex 10
2.3.6 Daitch-Mokotoff Soundex 11

3 Návrh a implementace **13**
3.1 Výčtové typy 13
3.2 Datové struktury 13
3.3 Modul pro porovnávání řetězců 14
3.4 Modul pro tvorbu tříd jmen 15
3.5 Modul pro spojování genealogických záznamů 16

4 Experimenty **18**
4.1 Tvorba tříd 18
4.1.1 Způsob porovnání tříd 19
4.1.2 Postup hledání počtu chyb pro první typ chyby 19
4.1.3 Postup hledání počtu chyb pro druhý typ chyby 20
4.1.4 Experiment č. 1 20
4.1.5 Experiment č. 2 22
4.1.6 Experiment č. 3 23
4.1.7 Experiment č. 4 24
4.2 Spojování záznamů 24
4.2.1 Párování se záznamy o narození rodičů 25
4.2.2 Párování záznamů o narození dětí 26
4.2.3 Problematika vdov 27
4.2.4 Experiment č. 5 27
4.2.5 Experiment č. 6 29

4.2.6 Experiment č. 7 30
4.2.7 Další experimenty a měření 31
4.3 Celkové zhodnocení 32
4.4 Návrhy do budoucna 32

5 Závěr **34**

Literatura **35**

A Obsah CD **36**

B Zdrojová data **37**

Kapitola 1

Úvod

V současné době je velmi populárním koníčkem lidí sestavování si rodokmenů svých rodin. Zjišťují si tak z matrik, kdo byli jejich předci, kde žili a kde žijí nějací jejich vzdálení příbuzní v současnosti. Spojování genealogických záznamů však neslouží pouze jako zábava pro lidi, ale slouží i k vědeckým účelům zejména v sociologii. Lze tak pomocí toho zjišťovat rozdělení společnosti v minulosti, zkoumat interakce mezi těmito skupinami a vyvozovat z toho poznatky do současnosti.

Tato práce má za cíl prostudovat metody slučování dat v historických dokumentech, seznámit se způsoby zápisu těchto dat a identifikovat problémy při identifikaci rolí na dodané sadě dat.

V kapitole 2 si proto rozebereme nejdříve obecný problém slučování dat z různých souborů záznamů. Seznámíme se s různými metodami pro porovnávání řetězců, jelikož v genealogický záznamech se setkáváme především s řetězci (jména a příjmení osob, jejich povolání, bydliště, apod.). A také si probereme rozdělení pravděpodobnosti spojité náhodné velikosti, které budeme například využívat v případě, když si budeme klást otázku, s jakou pravděpodobností se daný člověk mohl dožít daného věku tj., že se budeme snažit spojit záznam o narození člověka s nějakým záznamem o úmrtí člověka.

V kapitole 3 je pak popsán návrh a implementace základních modulů programu, který slouží pro experimentování na dodané sadě dat. Na tuto kapitolu pak navazuje delší kapitola 4, ve které jsou popsány jednotlivé experimenty, které pomáhaly zjišťovat problémy při identifikaci rolí. První část experimentů se týká třídění různých tvarů jmen do tříd, které reprezentuje normovaný tvar jména. Druhá část experimentů se pak týká již samotného slučování záznamů. Součástí těchto experimentů je také jejich zhodnocení, které je důležité pro další pokračování v této oblasti.

V závěrečné kapitole 5 je poté shrnuto vše, co lze zjistit přečtením této práce a dále pak základní shrnutí toho, co by bylo vhodné doplnit při následujícím pokračování v této práci.

Kapitola 2

Teorie

Tato práce se zabývá spojováním záznamů v genealogických datech, v této kapitole bude proto provedeno stručné seznámení s jednotlivými technikami, které se používají k pravděpodobnostnímu spojování záznamů.

Nejdříve bude probrána teorie Fellegi-Sunter[4] pro spojování záznamů ze dvou datových souborů. Tato teorie z roku 1969 je formalizovaným základem moderní teorie spojování záznamů, ze které se vychází až do současnosti.

V následujících sekcí pak budou rozebrány techniky podle toho, jakého se týkají typu dat. S těmito technikami se pak bude přímo pracovat v praktické části. U genealogický dat se to týká především srovnávání dat, jelikož u každého záznamu máme vždy nějaké datum narození, oddaní nebo úmrtí a program musí na základě rozdělení pravděpodobnosti určit, jaká je pravděpodobnost, že například záznam o narození a záznam o úmrti se týkají té stejné osoby. Druhým důležitým typem dat, který se vyskytuje ve zdrojových záznamech, jsou pak řetězce a to zejména jména a příjmení, které se u každého záznamu vyskytují vždy několikrát (jméno a příjmení dotčené osoby, jméno jejího otce, matky, apod.). V této části pak bude probráno několik metod, kterými lze porovnávat řetězce a budou se u nich posuzovat jejich možné nevýhody pro tento problém. Tyto metody budou poté v praktické části přímo testovány na reálných datech.

2.1 Teorie Fellegi-Sunter

Roku 1959 položil H. B. Newcombe základy pravděpodobnostního moderního spojování záznamů. O deset let později je pak formalizovali I. Fellegi a A. Sunter a jejich práce *A Theory For Record Linkage* je až dodnes matematickým základem mnoha aplikací pro spojování záznamu. Tato část slouží tedy k seznámení s těmito matematickými základy na příkladu spojování dvou souborů se záznamy.

Mějme dva soubory záznamů (též populace) A a B. Jednotlivé jejich záznamy označme jako a a b. Zároveň předpokládáme, že některé záznamy jsou společné pro oba soubory. Pak máme následující množinu uspořádaných dvojic:

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

Tato množina je sjednocením množin M a U, kde množina M obsahuje sobě odpovídající záznamy, zatímco množina U navzájem neodpovídající záznamy z obou souborů záznamů:

$$M = \{(a, b) : a = b, a \in A, b \in B\}$$

$$U = \big\{(a,b) : a \neq b, a \in A, b \in B\big\}$$

Dále označme záznamy korespondující si prvky množin A a B jako $\alpha(a)$ a $\beta(b)$. Pak je definován porovnávací vektor γ následovně:

$$\gamma\big[\alpha(a),\beta(b)\big] \equiv \Big\{\gamma^1\big[\alpha(a),\beta(b)\big],\gamma^2\big[\alpha(a),\beta(b)\big],\cdots,\gamma^K\big[\alpha(a),\beta(b)\big]\Big\}$$

Každý prvek γ^i pro $i = 1,\ldots,K$ představuje specifické porovnání prvků obou množin. Například γ^1 může představovat porovnání příjmení z množiny A s příjmením z množiny B a γ^2 zase může představovat porovnání adresy prvků s obou množin. Každé z těchto porovnání nám pak dává určitou hodnotu toho, jak si záznamy odpovídají.

Jako A_1 pak značíme taková rozhodnutí, kdy je mezi záznamy a a b navázána vazba na základě porovnávacího vektoru γ (tj., že záznamy považujeme za shodné). Jako A_2 značíme rozhodnutí, že mezi záznamy a, b je možná vazba, a jako A_3 značíme rozhodnutí, že mezi záznamy a, b není vazba. Pravidlo vazby L pak definujeme jako zobrazení z Γ (srovnávacího prostoru) na množinu náhodných rozhodovacích funkcí $D = \big\{d(\gamma)\big\}$, kde:

$$d(\gamma) = \big\{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\big\}; \gamma \in \Gamma$$

$$\sum_{i=1}^{3} P(A_i|\gamma) = 1$$

Dále mohou nastávat dva druhy chyb. První chybou je to, že dva záznamy jsou označeny za shodné, i když ve skutečnosti nejsou. Tato chyba má pravděpodobnost $P(A_1|U)$. Druhým typem chyby je to, že jsou naopak označeny dva záznamy za neshodné, i když ve skutečnosti jsou shodné. Tato chyba má pravděpodobnost $P(A_3|M)$.

Fellegi a Sunter definovali pravidlo vazby L_0 spolu se spojenými rozhodnutími A_1, A_2 a A_3 jako optimální dle následujícího teorému.

Teorém (Fellegi-Sunter 1969). Nechť L' je pravidlo vazby spolu se spojenými rozhodnutími A'_1, A'_2 a A'_3 takové, že má stejné pravděpodobnosti chyby $P(A'_3|M) = P(A_3|M)$ a $P(A'_1|U) = P(A_1|U)$ jako L_0. Pak L_0 je optimální, pokud $P(A'_2|U) \leq P(A_2|U)$ a $P(A'_2|M) \leq P(A_2|M)$.

Pro tuto práci z toho plyne následující:

- Bude třeba si zvolit způsob porovnání mezi jednotlivými atributy našich záznamů (mezi příjmeními, mezi daty, apod.) a z těchto porovnání pak sestavit celkový porovnávací vektor γ.
- Pro výsledný porovnávací vektor γ pak bude potřeba si nastavit hranice, kdy budeme brát, že mají záznamy mezi sebou vazbu, kdy jí mohou mít a kdy spolu vůbec nesouvisí.
- Tyto hranice a jednotlivé porovnávací vektory optimalizovat tak, abychom dostali vazbu jen mezi skutečně odpovídajícími záznamy a naopak, aby nám neunikla nějaká vazba u záznamů, které jsme označily za záznamy bez společné vazby.

2.2 Srovnání záznamů dle data záznamu

Pokud jsou porovnávány mezi sebou záznamy ze tří různých druhů matrik (tj. z matrik narozených, oddaných a zemřelých), je důležité si správně nastavit rozdělení pravděpodobnosti, aby pomocí něj šlo určit, s jakou pravděpodobností si dva záznamy odpovídají dle rozdílu dat, která se u jednotlivých záznamů vždy nacházejí. Například pokud je hledán k záznamu o narození z roku 1820 záznam o tom, kdy se daný člověk nechal oddat, tak nemůže být nalezena shoda v záznamech v matrice oddaných před rokem 1820.

V této sekci bude tedy nejdříve rozebráno obecné rozdělení pravděpodobnosti spojité náhodné veličiny a pak bude detailněji ukázán příklad jednoho takového rozdělení pravděpodobnosti (Normální rozdělení). Z tohoto rozdělení se pak částečně vychází v praktické části práce, kdy ale pro spojitou veličinu čas se pracuje vždy s intervalem daného roku. Proberou se zde však také stručně historické souvislosti, které mohou vnést do ideálních podmínek poměrně významné chyby, a je tedy potřeba na ně myslet a vzít je v potaz při výsledcích a případně je zahrnout už i do návrhu rozdělení pravděpodobnosti, a vytvořit tal pomocí nich jiné, které bude daným podmínkám více vyhovovat.

2.2.1 Rozdělení pravděpodobnosti spojité náhodné veličiny

Rozdělení pravděpodobnosti náhodné veličiny [3] udává, jaká je pravděpodobnost, že náhodná veličina nabývá zrovna dané hodnotě. Rozdělení pravděpodobnosti pro spojitou náhodnou veličinu se tedy dostane, pokud každému intervalu hodnot spojité náhodné veličiny je přiřazena pravděpodobnost.

Každé rozdělení spojité náhodné veličiny lze popsat pomocí funkce hustoty pravděpodobnosti $f(x)$ a distribuční funkci $F(x)$. Pro funkci hustoty pravděpodobnosti $f(x)$ platí následující rovnost (Ω je definiční obor veličiny X):

$$\int_{\Omega} f(x)dx = 1$$

Důležitým vztahem je pak vztah pro výpočet pravděpodobnosti, že náhodná veličina X nabývá hodnotu z intervalu $\langle x_1, x_2 \rangle$ (z tohoto vztahu také plyne, že pravděpodobnost toho, že spojitá náhodná veličina nabýva přesně dané hodnoty, je nulová):

$$P[x_1 \leq X \leq x_2] = \int_{x_1}^{x_2} f(x)dx$$

Pro distribuční funkci $F(x)$ platí vztah $F(x) = P(X \leq x)$ tj., je definována jako pravděpodobnost toho, že realizace náhodné veličiny X nepřekročí x. Pro tuto funkci platí, že v $-\infty$ je nulová, v $-\infty$ nabývá hodnoty 1 a je neklesající. Mezi funkcí hustoty pravděpodobnosti a distribuční funkcí platí následující vztah:

$$F(X) = \int_{-\infty}^{x} f(t)dt$$

2.2.2 Normální rozdělení

Mezi nejznámější a nejdůležitější normální rozdělení pravděpodobnosti patří normální rozdělení (někdy též označováno jako Gaussovo rozdělení). Toto rozdělení dobře modeluje náhodné děje vyskytující se v přírodě. Typickým příkladem je rozdělení populace podle IQ nebo výšky. Toho lze tedy částečně využít i pro problém spojování záznamů, kdy tímto

způsobem lze modelovat pravděpodobnost dožití (pokud se bude zvlášť uvažovat vysokou kojeneckou úmrtnost) nebo pravděpobnost toho, jakého věku byli ženich a nevěsta při první svatbě.

Funkce hustoty pravděpodobnosti $f(x)$ je pro normální rozdělení definována pomocí následujícího vztahu:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}},$$

kde μ je střední hodnota a σ^2 je rozptyl.

Pro distribuční funkci $F(x)$ pak lze dostat s využitím obecného vztahu mezi funkcí hustoty pravděpodobnosti a distribuční funkcí následující rovnost:

$$F(X) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt$$

2.2.3 Historické souvislosti

Pokud jsou porovnávány jakékoliv matriční záznamy, je důležité vědět, jestli se v té době nepromítly na daném místě nějaké přelomové historické události, které mohli ovlivnit počet nově narozených, úmrtnost nebo počet nově založených svazků. Mezi takové události mohou patřit světové války, které zasáhly skoro každé území, ale také různé epidemie moru, které mohly v daném místě. S těmito dodatečnými informacemi je tedy důležité počítat a brát je na zřetel při určování pravděpodobnosti shody dvou záznamů.

Tato práce se zabývá především obdobím mezi lety 1686 až 1850 ve venkovském prostředí jižní Moravy. Poslední a jediná v tomto období zasáhla velká epidemie moru Moravu v roce 1714 [5]. Tím, že se ale pohybujeme ve venkovském prostředí, tak zde nepůsobila tolik jako ve městě (například v Praze během této epidemie zemřela čtvrtina obyvatel).

Další důležitou historickou událostí v této době jsou napoleonské války, které probíhaly na začátku 19. století. Ty způsobily na našem území tzv. hladová léta [2], takže v tomto období narostla především kojenecká úmrtnost, kdy v prvním měsíci života umírala až polovina všech novorozenců.

Kojenecká úmrtnost a úmrtnost dětí do 5 let byla vysoká v celém tomto období, proto je důležité s nimi počítat při spojování záznamů. Průměrně se odhaduje kojenecká úmrtnost na začátku 18. století před začátkem průmyslové revoluce kojenecká úmrtnost zhruba kolem 33 % a do 5 let umírala skoro polovina dětí. Po začátku průmyslové revoluce se tyto údaje začínají zmenšovat a v roce 1870 se kojenecká úmrtnost pohybovala okolo 25 %. [1]

Vysoké údaje u úmrtnosti malých dětí pak také zkreslují údaje o střední délce života v této době. Do ní se totiž zahrnují i tato úmrtí, a věkový údaj tak může být i o 50 % menší než ve skutečnost. Tzn., že osoby, které se dožily dospělosti, se průměrně mohly dožít i věku kolem 60 let a ne pouze 30 let, což bývá udáváno jako střední délka života p5ed pr;myslovou revolucí.

Při modelování rozdělování pravděpodobnosti lze tedy uvažovat a zahrnout všechny tyto údaje.

2.3 Metody pro porovnávání řetězců

Při porovnávání dvou záznamů z matriky, u kterých je snahou najít shodu, je důležitou součástí srovnání jmen a příjmení osob, které jsou u daných záznamů uvedeny. V této rozsáhlé sekci se proto rozeberou metriky, které se používají k měření podobnosti dvou řetězců. Zkusí se u nich odhadnout, zda jsou vhodné při řešení našeho problému spojování záznamů v genealogických datech, a v kapitole 4 pak provedeme experiment a vyhodnocení metod na reálných datech.

Nejdříve budou projity metriky podobnosti, které vycházejí z úprav původního řetězce na jiný řetězec. Mezi tyto metriky patří Hammingova vzdálenost, která patří mezi implementačně nejjednodušší, dále pak Levenshteinova vzdálenost, která bývá nazývána též jako editační vzdálenost, Jarova vzdálenost a Jaro-Winklerova vzdálenost, která je rozšířením Jarovy vzdálenosti.

Druhým typem metrik, který zde bude detailněji rozebrán, jsou metriky založené na fonetických shodách. Mezi nejznámější z nich patří Soundex a Daitch-Mokotoff Soundex, který byl vytvořen speciálně pro východoevropské příjmení. Tyto metriky by mohly být pro problém řešený v této práci vhodné, jelikož jména i příjmení se vyvíjela mj. v závislosti na jejich výslovnosti.

2.3.1 Hammingova vzdálenost

Hammingova vzdálenost dvou řetězců je rovna počtu pozic, na kterých se v řetězcích nachází odlišné znaky. Například pro příjmení Godinict a Godenich by se tedy tato vzdálenost vyhodnotila jako 2, jelikož se řetězce liší na čtvrté a osmé pozici. Výhodou této metody je jednodušší výpočet, a tedy i snadná implementace.

První nevýhodou této metody je, že lze porovnávat pouze řetězce stejné délky. Této nevýhody se lze zbavit například tím, že spočítáme vzdálenost na prvních x pozicích, kde x je délka kratšího řetězce, a k této vzdálenosti pak přičteme rozdíl délek obou řetězců. Pro příjmení Groh a Grahovina se tak dostane vzdálenost 6.

Druhou nevýhodou této metody je to, že neposuzuje vzdálenost vzhledem k délce obou řetězců. Proto například při srovnání příjmení Anderschitz s Anderschitzin se dostává vzdálenost 2, což je stejná vzdálenost jako při srovnání příjmení Belan a Bihan, přitom je vidět, že shoda by u první dvojice měla být vyšší. Tento problém lze trochu zlepšit tím, že by se vzdálenost například podělila délkou delšího řetězce.

Dalším problémem, který tato metoda neřeší, je pak váha pozice, na které je nalezena odlišnost ve znacích. Tj. například v případě, že se dvě příjmení liší na první pozici, tak by pravděpodobnost jejich shody měla být menší než, když se liší na poslední pozici, která je nachýlnější na různé koncovky.

2.3.2 Levenshteinova vzdálenost

Levenshteinova vzdálenost [6] dvou řetězců je rovna minimálnímu počtu operací, které jsou potřeba k úpravě jednoho řetězce na druhý. Tato metrika pracuje se třemi typy operací:

- Vkládání znaku - přidání znaku na začátek, dovnitř nebo na konec řetězce. Například pokud máme příjmení Babitsh a Babitsch, stačí vložit znak c do prvního příjmení a dostáváme již shodu.
- Smazaní znaku - odebraní libovolné znaku z řetězce. Například opět pro příjmení Babitsh a Babitsch stačí smazat znak c z příjmení Babitsch a dostaváme opět shodu.

- Záměna znaků - nahrazení libovolné znaku jiným znakem. Například pro příjmení Arbes a Orbes pak stačí nahradit první znak A v prvním příjmení za O a dostáváme opět shodu.

Pro výpočet Levenshteinovy vzdálenosti existuje algoritmus Wagner-Fisher. Principem algoritmu je postupné vypočítávání hodnot v matici o rozměrech $n+1$ a $m+1$, kde n je délka prvního řetězce a m je délka druhého řetězce. Na začátku se do prvního řádku matice vloží hodnoty 0 až $n+1$ a do prvního sloupce pak hodnoty 0 až $m+1$. Následně se zahájí po řádcích postupné procházení celé matice. Hodnota v aktuálním poli matice $mat[i][j]$ se vypočítá dle následujícího vzorce (s nabývá hodnoty 1, pokud se znaky na pozici i a j liší, jinak nabývá hodnoty 0):

$$mat[i][j] = min(mat[i-1][j]+1, mat[i][j-1]+1, mat[i-1][j-1]+s)$$

Výsledná vzdálenost se pak nachází na pozici $mat[n][m]$.

V tabulce 2.1 lze vidět jednotlivé hodnoty v matici při porovnávání příjmení Jandreschitz a Jankovitz. Celková Levenshteinova vzdálenost těchto dvou příjmení je 6.

Tabulka 2.1: Matice po výpočtu Levenshteinovy vzdálenosti

		J	A	N	D	R	E	S	C	H	I	T	Z
	0	1	2	3	4	5	6	7	8	9	10	11	12
J	1	**0**	1	2	3	4	5	6	7	8	9	10	11
A	2	1	**0**	1	2	3	4	5	6	7	8	9	10
N	3	2	1	**0**	1	2	3	4	5	6	7	8	9
K	4	3	2	1	**1**	2	3	4	5	6	7	8	9
O	5	4	3	2	2	**2**	3	4	5	6	7	8	9
V	6	5	4	3	3	3	**3**	**4**	**5**	6	7	8	9
I	7	6	5	4	4	4	4	4	5	**6**	**6**	7	8
T	8	7	6	5	5	5	5	5	5	6	7	**6**	7
Z	9	8	7	6	6	6	6	6	6	6	7	7	**6**

Nevýhodou Levenshteinovy vzdálenosti je podobně jako u Hammingovy vzdálenosti to, že nebere ohled na délky porovnávaných řetězců. To lze ale opět napravit tím, že výsledná vzdálenost se bude normovat podle délky delšího řetězce. Dále také tato metoda nerozlišuje, zda k odlišnostem dochází na začátku nebo na konci řetězce. Existují ovšem úpravy této metody, které přiřazují hodnotě s při záměně písmen různé váhy než jen 1 nebo 0. Řeší se tím, tak například překlepy na klávesnici, kdy pro písmena, která jsou vedle sebe na klávesnici se nastaví menší váha než pro vzdálenější.

2.3.3 Jarova vzdálenost

Jarova vzdálenost d_j [7] dvou řetězců je dána vztahem:

$$d_j = \begin{cases} \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{pro } m \neq 0 \\ 0 & \text{pro } m = 0 \end{cases}$$

Kde m je počet odpovídajících si znaků, $|s_i|$ jsou délky obou řetězců a t je polovina počtu transpozic. Dva znaky si odpovídají, pokud jsou stejné a nacházejí se na pozicích,

jejichž rozdíl nepřesahuje hodnotu:

$$\left\lfloor \frac{max\big(|s_1|, |s_2|\big)}{2} \right\rfloor - 1$$

Jelikož počet odpovídajících znaků nemůže být větší než délka kratšího řetězce, tak jednotlivé zlomky $\frac{m}{|s_1|}$, $\frac{m}{|s_2|}$ a $\frac{m-t}{m}$ jsou menší než 1 (a zároveň i větší než 0), takže hodnota $d_j \in \langle 0, 1\rangle$. Platí, že čím je výsledná hodnota blíže k 1, tak tím je pravděpodobnost shody větší.

Výpočet vzdálenosti si ukážeme na příkladu dvou příjmení Hudek a Huleck. Pak vzdálenost pozic odpovídajících si znaků nesmí přesahovat hodnotu 2 ($\lfloor \frac{max(5,6)}{2} \rfloor - 1$). Hodnota m je tedy v našem případě rovna 4, jelikož znaky 'H', 'u', 'e' se nachází na stejné pozici a znak 'k' se liší pouze o 1 pozici. Hodnota t je nulová, jelikož v obou řetězcích jsou odpovídající si znaky ve stejném pořadí. S těmito znalostmi můžeme již dopočítat celkovou vzdálenost porovnávaných řetězců:

$$d_j = \frac{1}{3}\Big(\frac{4}{5} + \frac{4}{6} + \frac{4-0}{4}\Big) = \frac{37}{45} \doteq 0.822$$

Tato metoda tedy narozdíl od předchozích pracuje i s délkou řetězců, stále však nepracuje s váhou pozic tedy, že pokud dojde k záměně na prvních pozicích řetězců, tak bude shoda jiná, než když dojde k záměně na koncových místech v řetězci.

2.3.4 Jaro-Winklerova vzdálenost

Jaro-Winklerova vzdálenost [7] je rozšířením Jarovy vzdálenosti. K ní přidává právě důraz na společný prefix obou řetězců, čehož se využívá právě při porovnávání křestních jmen a příjmení. Jaro-Winklerova vzdálenost d_w lze vypočítat podle následujícího vzorce, kde d_j je Jarova vzdálenost vypočítaná pro porovnávané řetězce, l je délka společného prefixu řetězců (nejvýše však 4) a p je škálovací faktor, který přidává váhu prvním l znakům (hodnota škalovacího faktoru nesmí přesáhnout 0.25, aby nebyla výsledná hodnota d_w větší než 1, zpravidla se používá hodnota 0.1):

$$d_w = d_j + lp(1 - d_j)$$

Například pro dvojici příjmení Hudek a Huleck, pro kterou už máme vypočítanou Jarovu vzdálenost, dostáváme následující výsledek (společný prefix má délku 2 a pro škálovací faktor volíme standartní hodnotu 0.1):

$$d_w = \frac{37}{45} + 2 \cdot 0.1 \cdot (1 - \frac{37}{45}) = \frac{193}{225} \doteq 0.858$$

Tato metoda však nepřidává nic, pokud se porovnávají například příjmení Paransitz a Baranchitz, které se liší již na první pozici. Přitom je vidět, že tato příjmení jsou si příbuzná, zejména po fonetické stránce.

2.3.5 Soundex

Soundex [8] je prvním fonetickým algoritmem, který je v této práci rozebírán a také je to jeden z prvně vytvořených fonetických algoritmů. V průběhu dalších let pak bylo vytvořeno několik algoritmů na podobném principu, které se snaží vylepšit problémy původního algoritmu.

Principem Soundexu je rozdělení souhlásek do několika skupin v závislosti na jejich výslovnosti. Jejich rozdělení dle původního Soundexu je znázorněno v tabulce 2.2. Každý řetězec se pak překládá na jiný řetězec dle těchto skupin, přičemž první písmeno se zachovává a samohlásky a některé souhlásky, které nejsou v tabulce se ignorují. Výsledný řetězec je poté zkrácen na 4 znaky (případně doplněn na 4 znaky odpovídajícím počtem nul).

S těmito informacemi lze dopočítat i maximální množství řetězců, které takto lze vytvořit (výpočet je prováděn pro anglickou abecedou, která má 26 znaků):

$$26 \cdot (6 \cdot 6 \cdot 6 + 6 \cdot 6 + 6 + 1) = 6734$$

Tabulka 2.2: Rozdělení souhlásek podle původního algoritmu

Skupina	Souhlásky
1	B, P, F, V
2	C, S, K, G, J, Q, X, Z
3	D, T
4	L
5	M, N
6	R

Tímto způsobem lze dostat například shodu pro dvojici příjmení Hoffpauer, Hofvbauer, které se obě překódují do tvaru H111. Podobně však dostáváme shodu i pro dvojici příjmení Bobek a Baveja, které se překódují do tvaru B120. Je ale patrno, že u první dvojice příjmení by měla být pravděpodobnost shody větší než u druhé dvojice. Tento problém se snaží vyřešit rozšířená verze algoritmu Soundex, kdy bylo vytvořeno pro souhlásky více skupin, což je vidět v tabulce 2.3. Také už není řetězec zarovnáván pouze na 4 znaky, ale jeho délka je proměnlivá. Počet řetězců vytvořených tímto způsobem je, tak podstatně větší, takže šance, že se k sobě napárují rozdílná příjmení je menší.

Tabulka 2.3: Rozdělení souhlásek podle upraveného algoritmu

Skupina	Souhlásky
1	B, P
2	F, V
3	C, K, S
4	G, J
5	Q, X, Z
6	D, T
7	L
8	M, N
9	R

Ani takto upravený algoritmus Soundex se však neumí vypořádat s dvojicí příjmení, která se liší například jen pouze v prvním písmenu, jelikož to je vždy v kódu zachováno.

2.3.6 Daitch-Mokotoff Soundex

Existuje mnoho fonetických algoritmů, které jsou specializované na nějaký jazyk (nejčastěji na anglický jazyk a jeho různé varianty). Zde je představen algoritmus Daitch-Mokotoff

Soundex [8], který se specializuje na východoevropské jazyky, což by mohlo částečně vyhovovat souboru zdrojových dat této práce, ve kterém se nacházejí převážně jména a příjmení, které pocházejí z příbuzné skupiny Jihoslovanských jazyků.

Tento algoritmus používá ke kódování číslice podobně jako Soundex, ale na rozdíl od něj se kóduje i první písmeno řetězce a také se kódují i celé skupiny znaků. Rozdíl je také v tom, že o skupině rozhoduje i to, kde se v řetězci daný znak nebo skupina nachází. V tabulce 2.4 jsou přehledně popsány tyto konverze.

Tabulka 2.4: Kódování podle Daitch-Mokotoff Soundex

Kombinace	Na začátku	Po samohlásce	Jinde
AI, AJ, AY, EI, EY, EJ, OI, OJ, OY, UI, UJ, UY	0	1	
AU	0	7	
IA, IE, IO, IU	1		
EU	1	1	
A, UE, E, I, O, U, Y	0		
J	1	1	1
SCHTSCH, SCHTSH, SCHTCH, SHTCH, SHCH, SHTSH, STCH, STSCH, STRZ, STRS, STSH, SZCZ, SZCS	2	4	4
SHT, SCHT, SCHD, ST, SZT, SHD, SZD, SD	2	43	43
CSZ, CZS, CS, CZ, DRZ, DRS, DSH, DS, DZH, DZS, DZ, TRZ, TRS, TRCH, TSH, TTSZ, TTZ, TZS, TSZ, SZ, TTCH, TCH, TTSCH, ZSCH, ZHSH, SCH, SH, TTS, TC, TS, TZ, ZH, ZS	4	4	4
SC	2	4	4
DT, D, TH, T	3	3	3
CHS, KS, X	5	54	54
S, Z	4	4	4
CH, CK, C, G, KH, K, Q	5	5	5
MN, NM		66	66
M, N	6	6	6
FB, B, PH, PF, F, P, V, W	7	7	7
H	5	5	
L	8	8	8
R	9	9	9

Celkem tento algoritmus poskytuje okolo 600 000 různých hodnot kódů. Například pro příjmení Parantshits a Barantchitz se dostane stejný kód 79644.

Kapitola 3

Návrh a implementace

Tato kapitola pojednává o návrhu a implementaci programu, který dostává na vstupu zdrojová data z matrik narozených, oddaných a zemřelých. První část programu pak s těmito daty umožňuje provádět různé experimenty, a na základě jejich výsledků tak získat lepší podklady pro druhou část programu, která se zabývá už spojováním dat a vytvořením základních rodinných stromů.

Celý program byl tedy navrhnut tak, že je rozdělen do několika modulů. Prvním z nich je modul pro samotné načtení dat ze zdrojových souborů. Tento modul se využívá pro každou část programu. Jako další byl navrhnut modul pro základní porovnávání dvou řetězců. Třetím modulem je pak modul pro tvorbu tříd jmen. Třída jmen je taková skupina různých tvarů jmen lidí, které jsou programem při spojování záznamů považovány za sobě odpovídající si tvary. Podrobněji je tvorba tříd jmen rozebírána v následující kapitole v sekci 4.1. V této kapitole jsou zmíněny funkce, které jsou implementovány pro jejich vytváření, a funkce, které slouží pro následovné ověřování jejich správnosti. Tento část tedy slouží zejména pro první část programu, ale jeho výstupy jsou použity pro druhou část.

Posledním navrženým modulem je modul pro spojování genealogických modelů. Používá se tedy primárně až pro druhou část programu a obsahuje funkce na porovnání záznamů, vypsání základních rodin a zobrazení statistik.

3.1 Výčtové typy

V souboru `distances.hpp` jsou definovány 2 výčtové typy. Prvním výčtovým typem je `Sex`, který obsahuje tři členy s identifikátory `Man`, `Woman` a `Other`. Tento výčtový typ má hlavní uplatnění při určování toho, jestli je dané jméno pro ženské nebo mužské pohlaví a také je to jeden z atributů pro záznam o narození dítěte, který není určen ve zdrojových datech.

Druhým výčtovým typem je pak `TypeName`, který má dva členy s identifikátory `Name` a `Surname`. Tento typ je využíván jako pomocný pro určování, jestli některé funkce pracující aktuálně se jmény nebo s příjmeními a podle toho mohou pracovat trochu odlišně (pracovat s jinými asociativními poli, apod.).

3.2 Datové struktury

Program obsahuje celkem 6 heterogenních datových struktur, přičemž nejdůležitějšími třemi základními strukturami jsou ty, které slouží zejména pro uložení zdrojových dat ze tří typů matričních záznamů. Všechny tyto struktury jsou definovány v souboru `main.hpp`.

První z těchto základních struktur je struktura `sBirth`, která obsahuje tedy především členy, do kterých se ukládají hodnoty ze zdrojových dat z matriky narozených. Mezi tyto údaje patří jméno a příjmení dítěte či rodičů, datum narození nebo adresa. Dalším důležitým členem je ukazatel na další prvek této struktury. Tento člen tak umožňuje zjednodušený průchod přes všechny záznamy daného typu, čehož se využívá v mnoha funkcích. Dalším členem je pohlaví dítěte, jehož datovým typem je již výše zmiňovaný výčtový typ `Sex`, hodnota tohoto členu se určuje až za běhu programu. Dále obsahuje struktura několik méně důležitých pomocných členů, které se používají při různých experimentech při spojování záznamů. Posledními významnými členy jsou pak prvky datového typu ukazatel na strukturu `sFamily`, která je rozebírána dále v textu. Těchto ukazatelů je více, protože jeden záznam může figurovat ve více rodinách (v jedné jako narození dítě a ve více jako narození rodiče). Tyto členy se využívají až po spojení záznamů pro jednodušší průchod mezi jednotlivými základními rodinami.

Druhou základní strukturou je struktura `sMarriage`, která obsahuje především členy, do kterých se ukládají hodnoty ze zdrojových dat z matriky oddaných (jméno a příjmení ženicha s nevěstou, jména jejich rodičů, datum oddání či adresy). Další členem struktury je podobně jako pro strukturu sBirth člen, který je ukazatelem na další prvek struktury `sMarriage`. Posledním významnějším členem je pak prvek datového typu ukazatel na strukturu `sFamily`. Tento ukazatel je právě jeden, protože v každé rodině je pouze jedno oddání rodičů, to je rozdíl od struktury `sBirth`, která jich má více.

Poslední základní strukturou je struktura `sDeath`, která obsahuje členy, do kterých se ukládají hodnoty ze zdrojových dat z matriky zemřelých, a také člen next, jehož datový typem je ukazatel na strukturu `sDeath`. Tato struktura se však v dalších částech programu není příliš používána, jelikož zdrojových dat nebylo tolik, aby na nich šlo provádět větší experimenty.

Významnou strukturou, která už byla výše několikrát zmiňována, je struktura `sFamily`. Tato struktura obsahuje podobně jako předchozí struktury prvek `next`, který je datového typu ukazatel na strukturu `sFamily`. Dále pak obsahuje prvky, které jsou ukazateli na výše zmíněné struktury. V těchto ukazatelích je tak uložena informace, jaký záznam o oddání se vztahuje k této rodině, jaké záznamy o narození jsou pravděpodobné záznamy narození rodičů a které všechny záznamy o narození jsou pravděpodobně záznamy o narození dětí v této rodině. Pro každou tuto informaci existuje vždy i prvek datového typu double, který vyjadřuje váhu správnosti tohoto spárování daného záznamu s danou rodinou.

Zbývající dvě struktury `sDate` a `sAddress` jsou pouze pomocné struktury pro podrobné uložení jednotlivých dat o narození, oddání a úmrtí a o bydlišti.

Program dále obsahuje několik asociativních polí, které se používají pro uložení informace o podobnosti jmen a jejich počtu. Tyto pole jsou blíže rozebrány níže v části o modulu pro tvorbu tříd jmen, ve kterém se nejvíce uplatňují.

3.3 Modul pro porovnávání řetězců

Funkce pro porovnání dvou řetězců, se nacházejí v souboru `distances.hpp`. Konkrétně se jedná o funkce, jejichž názvy jsou následující: `hammingDistance`, `levensteinDistance`, `jarDistance` a `jarWinklerDistance`. Všechny tyto funkce mají stejný návratový typ double a vracejí vypočtenou vzdálenost. První dva parametry každé funkce jsou datové typu string a jde vždy o řetězce, které mají být mezi sebou porovnány. Funkce `jarWinklerDistance`

má pak ještě třetí parametr datového typu double, který je nepovinný s defaultní hodnotou 0.1 a jedná se o tzv. škálovací faktor. Algoritmy používané ve všech čtyř funkcí vychází z teorie probírané v předchozí kapitole v sekci 2.3.

Mimo tyto čtyři funkce se v souboru `distances.hpp` nachází i definice funkce distance. Tato funkce má stejný návratový typ jako předchozí funkce a taktéž její první dva parametry jsou pro porovnávané řetězce. Třetí parametr pak slouží pro určení, jaká funkce pro porovnání řetězců se má použít. Hlavním účelem této funkce je tedy přepínat mezi základními čtyřmi funkcemi, může však používat i jejich libovolné kombinace. Tato funkce se také proto většinou volá z ostatních funkcí, když je třeba porovnat nějaké dva řetězce

3.4 Modul pro tvorbu tříd jmen

V souboru `names.hpp` jsou implementovány všechny funkce, které se používají pro tvorbu tříd jmen a příjmení. Dále pak jsou zde i funkce, které slouží ke zjišťování počtu chyb, které vznikly chybným spárováním jmen a příjmení. Kromě těchto funkcí jsou zde definice i 7 asociativních polí, které se používají při jejich tvorbě a které slouží pro účely experimentů, o nichž se píše v další kapitole.

Nejdůležitějšími z těchto asociativních polí jsou `map<string, vector<string»names` a `map<string, vector<string»surnames`, ve kterých jsou uloženy jednotlivé skupiny a později třídy odpovídajících si jmen pro každé jméno ze zdrojových dat. Z jejich názvu je pak patrno, že jedno se používá pro jména, zatímco druhé pro příjmení. Klíčem těchto polí je vždy jméno a hodnotou je pak pole jmen, která patří do stejné skupiny jmen jako jméno v klíči.

Dalšími významnými asociativními poli jsou `map<string,map<string,int»normalize` a `map<string,map<string,int»namesClassNormCount` (resp. `surnamesClassNormCount`). První z těchto polí slouží pro zjištění počtu chyb prvního typu, zbylé dvě pak zjištění počtu pro druhý typ chyb. Pojem chyba prvního a druhého typu je rozebírána v části X. Klíčem tohoto pole je vždy normovaný tvar jména (resp. jméno zastupující třídu jmen) a hodnotou je pak další asociativní pole, kde klíčem je jméno zastupující třídu jmen (resp. normovaný tvar) a hodnotou je počet jmen dané třídy, která se normují na daný normovaný tvar.

Posledním asociativním pole je pak `map<string,Sex>sexNames`, ve kterém je uložena informace o tom, k jakému pohlaví se vztahuje, které jméno.

Funkce v modulu lze rozdělit na několik druhů podle toho, s jakými z předchozích asociativních polí pracují, a dále pak podle toho, jestli slouží pro naplnění polí nebo pro jejich vyhodnocení a vypsaní.

Pro práci s asociativními poli `names` a `surnames` je tedy implementováno celkem 8 funkcí. První dvě funkce s názvy `cmpName` a `cmpSurname` jsou klíčové pro tvorbu počátečních skupin příbuzných jmen. Prvním parametrem těchto funkcí je jméno, pro které se dohledává skupina příbuzných jmen. Dalšími dvěma parametry jsou poté ukazatel na první záznam o narození a na první záznam o oddání a ve funkci jsou pak využity k průchodu přes všechny záznamy. Zbylé parametry jsou nepovinné a slouží pro experimenty. První z těchto parametrů je hodnota maximální vzdálenosti, při jaké se berou dva řetězce ještě za shodné a zařazují se do stejné skupiny. Dalším parametrem je pak metrika podle, které se má daná vzdálenost dvou řetězců vypočítat, defaultně se používá pro výpočet Jaro-Winklerova vzdálenost. Poslední parametr, který je pouze u funkce `cmpName`, pak říká, zda se má používat kontrola na pohlaví jména. Další dvě funkce s názvy `joinNames` a `joinSurnames`, které

nemají žádné parametr, slouží pro přeměnu skupin jmen na třídy jmen. Algoritmy těchto funkcí spolu s příklady jsou rozepsány v následující kapitole v sekci 4.1.

Zbývající čtyři funkce slouží buď pro uložení získaných tříd jmen do souboru, nebo pro jejich načtení ze souboru zpět do asociativních polí. Toho se využívá zejména pro experimenty při spojování záznamů, kdy není potřeba pokaždé vytvářet třídy znovu. Každá z těchto má jeden parametr a tím je soubor pro uložení nebo načtení dat.

Pro vytváření polí `normalize`, `namesClassNormCount` a `surnamesClassNormCount` se využívají funkce s předponou insertInto a název daného pole. Naopak pro jejich zobrazení a vyhodnocení údajů se používají funkce s předponou show. Popis jejich algoritmů je rozepsán v následující kapitole v sekci 4.1.

Posledními významnějšími funkcemi tohoto modulu jsou funkce, jejichž názvy jsou `assignNameToSex` a `checkSexInClasses`. První funkce vytváří pole `sexNames`, zatímco druhá s tímto polem pracuje a upravuje na základě dat v něm třídy jmen tak, aby v jedné třídě nebyly jména různého pohlaví. Toho se využívá u experimentu č. 3 v následující kapitole, kde je i blíže popsán algoritmus těchto metod.

3.5 Modul pro spojování genealogických záznamů

Funkce tohoto modulu jsou implementovány v souboru `marriage.hpp`. Nejdůležitější funkcí z tohoto modulu je funkce s názvem `foundFamily`, ve které se porovnávají jednotlivé záznamy o oddání se všemi záznamy o narození. Tato funkce vrací ukazatel na prvně nalezenou rodinu, která je datového typu `sFamily`. Prvním a druhým parametrem této funkce jsou ukazatelé na první záznam o manželství a o narození. Dalším parametrem je informace o tom, jakou metrikou se má počítat dodatečná vzdálenost mezi dvěma řetězci. Toho se využívá ve chvíli, kdy neleží porovnávané řetězce ve stejné třídě jmen, ale mohlo by jít přesto o stejná jména, která pouze nebyla rozpoznána. Posledním parametrem je maximální vzdálenost, pro kterou se má tato dodatečnou vzdálenost brát ještě jako relevantní. Bližší popis algoritmu této funkce je popsán v následující kapitole u experimentů se spojováním záznamů.

Další významnou funkcí je funkce s názvem `getYearProbability`, která na vstupu dostává věkový rozdíl dvou záznamů a typ záznamu tzn., jestli se porovnává záznam o narození ženicha s jeho svatbou nebo záznam o narození dítěte se svatbou rodičů. Výstupem této funkce je pak pravděpodobnost toho, že bylo nalezeno spojení mezi těmito záznamy.

Pro porovnání dvou záznamů o oddání za účelem zjištění nalezení prvního manželství pro vdovu se používá funkce s názvem `findPreviousMarriage` (resp. pro normované tvary `findPreviousMarriageNorm`). Tyto funkce přijímají na vstupu jediný parametr, kterým by měl být ukazatel na první záznam o oddání. Uvnitř funkce se pak postupně ve dvou cyklech porovnávají jednotlivé záznamy o oddání (bližší popis porovnávání těchto záznamů je popsán v podsekci 4.2.3). Počítají se počty úspěšně spárovaných záznamů a zapisují se pak do textového souboru.

Důležitou funkcí je i funkce s názvem `cmpResult`, jejímž úkolem je porovnat výsledky získané spojováním za pomoci normovaných tvarů jmen s výsledky získanými při spojování pomocí nenormovaných tvarů. Funkce na vstupu dostává pouze ukazatel na první záznam o narození, aby se uvnitř funkce mohly v cyklu procházet všechny tyto záznamy. Při porovnání je pak už využito toho, že u každého záznamu o narození jsou uloženy ukazatele na

záznamy o oddání, ke kterým byly spárovány a to jak pro normované, tak i pro nenormované tvary jmen.

Tyto ukazatele na spárované záznamy o oddání se získávají ve třech funkcích s názvy `foundChildren`, `foundFather` a `foundMother`. V těchto funkcích se totiž procházejí rodiny (prvky datového typu `sFamily`), které byly získány ve funkci s názvem `foundFamily`, a to proto, že ve funkci `foundChildren` je nalezeno vždy několik možností pro záznam o narození otce, matky, či dítěte a zde se z nich vybírají ty nejpravděpodobnější a zároveň se u těchto nejpravděpodobnějších záznamů uloží odkaz na dané spárované manželství.

Zbývající funkce tohoto modulu procházejí nalezené výsledky, počítají statistiky, ukládají tyto statistiky do souborů nebo pouze slouží jako pomocné funkce výše zmíněným funkcím. Příkladem takových funkcí jsou funkce s názvem `countSameNumber`, `countSameVillage` nebo `countSameJob`, které slouží při měření počtu spárovaných záznamů, které mají pro osoby uvedenou shodnou adresu nebo práci.

Kapitola 4

Experimenty

V této části práce budou rozebrány experimenty, které byly prováděny na dodaných zdrojových datech. Tyto data jsou původem z matrik obce Frelichov (nyní Jevišovka) z časového rozmezí 1686 až 1850.

Problémem těchto dat je, že jména a příjmení se velmi často neshodují ve všech znacích, i když jde ve skutečnosti o stejné příjmení či jméno. Tento problém může být způsoben několika faktory. Těmito hlavními mohou být chybný zápis faráře do matriky, historický vývoj jmen a příjmení, zápis jmen a příjmení v různých jazycích (německy, česky) nebo chyba při získávání dat z matriky, přičemž je jedno, jestli tuto chybu udělá člověk, který přepisuje záznam nebo program, který je na tento přepis určen a špatně rozezná nějaké písmeno. Program vytvořený pro experimenty na spojování záznamů by se měl s tímto problémem umět vypořádat a vytvořit za tímto účelem třídy jmen a příjmení, ve kterých budou syntakticky podobné tvary. Tyto tvary se pak budou považovat za stejné, když na ně program narazí při spojování záznamů.

Ze zdrojových dat se využije zejména toho, že pro většinu jmen a příjmení byly dodány i jejich normované tvary, kterými už jsou tedy jména a příjmení do takových tříd rozděleny, a v experimentech je tak možnost vůči čemu porovnávat. Například pro jméno Jacobus je v datech dodaný normovaný tvar „JAKUB“.

Bližší popis jednotlivých údajů v dodaných datech je rozepsán v příloze B.

4.1 Tvorba tříd

Nejdříve je potřeba si vysvětlit způsob, jakým se budou třídy vytvářet. Základní princip je pro všechny experimenty stejný. Načtou se všechna zdrojová data a postupně se prochází jednotlivé záznamy. U každého záznamu se poté projdou všechna jména a příjmení a pro každé z nich se kontroluje, zda už nebyl stejný tvar nalezen dříve. Pokud byl, pokračuje se dál, pokud nebyl, začne probíhat proces porovnávání tvaru s tvary u všech záznamů. Toto porovnání se může lišit pro každý experiment a to zejména použitou metodou na porovnání řetězců, maximální vzdáleností, pro kterou se ještě berou pro danou metodu řetězce za shodné nebo dalšími zjištěnými zkušenostmi.

Tímto způsobem se vytvoří pro každý tvar jména a příjmení skupina, ve které jsou jiné tvary, jejichž maximální vzdálenost od daného tvaru je nějaké zvolené x. Například pro Jaro-Winklerovu metodu, kdy jsou brány za shodné všechny tvary se vzdáleností větší než 0.95, dostáváme pro příjmení tvaru „BABICH“ skupinu příjmení „BABITSCH, BABISCH, BABITCH, BABICH“. Ve skupině pro tvar příjmení „BABITH“ jsou zase příjmení

„BABITSCH, BABITSH, BABITCH, BABITH“. Zde si lze všimnout, že v obou skupinách se nachází tvary „BABITSCH a BABITCH“, ale samotná vzdálenost tvarů „BABICH“ a „BABITH“ je menší než 0.95, takže jejich možná shoda ještě nebyla potvrzena.

Dále se tedy provádí rozšiřování skupin o nové tvary na základě zjištění podobných těm jako u skupin v příkladu. Tedy do skupiny pro příjmení tvaru „BABICH“ se přidají všechny tvary příjmení, které jsou ve skupinách pro tvary „BABITSCH, BABISCH a BABITCH“. To vše probíhá do doby než je splněna tranzitivita shodnosti tvarů příjmení, tj. když ve skupině pro tvar A je tvar B a ve skupině pro tvar C je také tvar B, tak i tvar C je ve skupině pro tvar A a tvar A je ve skupině pro tvar C.

Tyto výsledné skupiny už jsou pak třídami. Například pro tvar „Babich“ jsou tedy za daných podmínek brány za stejné tvary „BABICH, BABITSCH, BABITSCHIN, BABITSH, BABISCH, BABITCH, BABITS a BABIS“. Neberou se ale už za stejné například tvary „BABITZ, BABITCZ, BABICZ“.

4.1.1 Způsob porovnání tříd

Dále je potřeba vysvětlit způsob, jakým se bude porovnávat, jestli byly třídy správně vytvořeny. Mohou nastat dvě chybové situace, kdy je spárování jmen a příjmení chybné a na základě počtu těchto špatných spárování se tedy může porovnat, jaký způsob je vhodnější.

První chybnou situací (chyba prvního typu) je, že do jedné třídy jsou přidělena jména, jejichž normované tvary si neodpovídají. Například může být v jedné třídě příjmení „JANDERSCHITZ“ a „JAESCHITZ“, jejichž Jaro-Winklerova vzdálenost je asi 0.933, což není mnoho, protože například pro dvojici „ADNRESCHITZ“ a „JENDRESCHUTZ“ je tato vzdálenost asi 0.801. Přitom normované tvary příjmení pro první dvojici jsou „ANDERSIC“ a „JAKSIC“, zatímco pro druhou dvojici je to jeden shodný tvar „ANDERSIC“.

Druhou chybnou situací (chyba druhého typu) je pak to, když dvě jména se stejným normovaným tvarem jsou přiděleny do dvou odlišných tříd. To může nastat právě pro již zmiňovanou dvojici „ADNRESCHITZ“ a „JENDRESCHUTZ“, jejichž vzdálenost je na tolik velká, že se vytvoří spíše dvě různé třídy.

Obecně tak lze vidět, že obě chyby jdou proti sobě a je potřeba nalézt nějaký kompromis. Pokud bude limit na vzdálenost dvou jmen vysoký (tj. nebudou se smět příliš lišit), bude vytvořeno více tříd (v extrémním případě pro každý tvar jména jedna třída), a bude tedy nastávat málo chyb jako v první situaci, ale více z druhé situace. Naopak pokud bude limit na vzdálenost dvou jmen nízký, bude existovat méně tříd, takže bude více chyb jako v první situaci, ale méně v té druhé.

4.1.2 Postup hledání počtu chyb pro první typ chyby

Cílem je získat výsledné asociativní pole, kde klíčem je normovaný tvar jména a hodnotou je další asociativní pole (dále se bude o tomto poli mluvit jako o vnořeném asociativním poli). Klíčem v tomto vnořeném asociativním poli jsou pak jména, pro která platí, že za každou třídu je zde pouze jedno jméno, a hodnotou v tomto poli je zde pak počet jmen, které se z dané třídy mapují na dané normované jméno. Za třídu, která se mapuje na normovaný tvar, pak prohlásíme tu s největším počtem jmen v daném vnořeném asociativním poli. Počtem chyb u jednoho normovaného tvaru je pak součet počtů u všech ostatních tříd.

Příkladem jednoho prvku z popsaného asociativního pole může být následující prvek. Klíčem je normovaný tvar „BABIC“. Ve vnořeném asociativním poli jsou pak „BABITSCH“ $\Rightarrow$ 242, „BABITZ“ $\Rightarrow$ 129, „BABITSHOVITZ“ $\Rightarrow$ 1 a „VABITSCH“ $\Rightarrow$ 1. „BABITSCH“

je reálný tvar ze záznamů a je to reprezentant třídy, do které patří ještě například příjmení „BABITCH, BABITSCHIN“, „BABITSH“, číslo 242 pak říká, že existuje celkem 242 těchto příjmení, pro které je normovaným tvarem slovo „BABIC“. Tato třída se také označí jako správná. Jako chybně nespárované se však označí třída s příjmením „BABITZ“ (v ní jsou i příjmení „BABICZ“ a „BABITCZ“) a třídy „BABITSHOVITZ“ a „BABITSCCHIN“, které obsahují pouze jeden tvar příjmení. Celkový chybný počet příjmení pro tento prvek je tak 131.

K získání tohoto asociativního pole je zapotřebí projít přes všechny záznamy o narození a o oddání. U každého záznamu se kontrolují pouze ta jména, pro která existuje i jejich normovaný tvar. Pokud ještě daný normovaný tvar nebyl nalezen, přidá se nový prvek do asociativního pole, kde klíčem je nový normovaný tvar a hodnotou je asociativní pole s jedním prvkem, kde klíčem je jméno a hodnotou je číslo 1. Pokud už daný normalizovaný tvar existuje, projdeme všechny dosavadní klíče ve vnořeném asociativním poli a hledáme, zda aktuální jméno nepatří do stejné třídy jako některý z klíčů. Pokud ano, zvedneme u něho počet o jedna. Pokud ne, založíme nový prvek s hodnotou 1 ve vnořeném asociativním poli.

4.1.3 Postup hledání počtu chyb pro druhý typ chyby

Cílem je získat podobné asociativní pole jako pro první typ chyby, akorát tentokrát je klíčem vždy jedno jméno reprezentující třídu, klíčem ve vnořeném asociativním poli je normovaný tvar, pro který platí, že alespoň jedno jméno z dané třídy má tento normovaný tvar, hodnotou je pak opět počet takovýchto případů. Za normovaný tvar, který se mapuje na danou třídu, pak prohlásíme ten s největší hodnotou. Počtem chyb u jedné třídy je pak součet počtů u všech ostatních normovaných tvarů, tedy podobně jako u prvního typu chyby.

Příkladem jednoho prvku z popsaného asociativního pole může být následující prvek. Klíčem je příjmení „ANDRESCHITZ“, který je zástupcem třídy, ve které jsou společně s ním například příjmení „ANDREASCHITZ“, „JANDERSCHITZ“, „JAJSCHITZ“ nebo „JAKSCHITZ“. Ve vnořeném asociativním poli jsou pak „ANDERSIC“ ⇒ 227 a „JAKSIC“ ⇒ 18. Normovaný tvar „ANDERSIC“ se tedy označí jako správný. Celkový chybný počet příjmení pro tento prvek je pak tedy 18.

Toto asociativního pole se získává podobně jako pro první typ chyby. Nejprve je zapotřebí projít přes všechny záznamy o narození a o oddání, kde se vyberou ty jména s normovaným tvarem. Pokud ještě neexistuje prvek, jehož klíčem je jméno ze stejné třídy jako aktuální, přidá se nový prvek do asociativního pole, kde klíčem je jméno a hodnotou je asociativní pole s jedním prvkem, kde klíčem je normovaný tvar jména a hodnotou je číslo 1. Pokud je nalezeno jméno ze stejné třídy, projdeme všechny dosavadní klíče ve vnořeném asociativním poli a hledáme, zda normovaný tvar aktuálního jména není shodný s některým klíčem. Pokud ano, zvedneme u něho hodnotu o jedna. Pokud ne, založíme nový prvek s hodnotou 1 ve vnořeném asociativním poli.

4.1.4 Experiment č. 1

Při prvním experimentu byla při počátečním třídění do tříd použita Jaro-Winklerova vzdálenost, kdy se postupně měnila požadovaná minimální vzdálenost od 0.9 do 1. Počty chybných spárování jmen jsou přehledně vidět v tabulce 4.1, kde v 1. sloupci je zvolená Jaro-Winklerova vzdálenost, ve 2. a 3. sloupci jsou počty správně (resp. chybně) roztříděných jmen a příjmení pro chybu prvního typu, ve 4. a 5. sloupci jsou počty správně (resp. chybně) roztříděných jmen pro chybu druhého typu, ve 6. a 7. sloupci jsou počty správně (resp.

chybně) roztříděných příjmení pro chybu druhého typu a v posledním sloupci je celkový počet chyb.

Tabulka 4.1: Výsledky experimentu č. 1

d_w	1. Sp.	1. Šp.	2. př. Sp.	2. př. Šp.	2. jm. Sp.	2. jm. Šp.	Celkem chyb
0.99	30445	17207	15313	0	32339	0	17207
0.98	30699	16953	15303	10	32333	6	15969
0.97	34397	13255	15255	58	32177	162	13475
0.96	39121	8531	15198	115	31139	1200	9846
0.95	40954	6698	15074	239	28104	4235	11172
0.94	42059	5593	14783	530	28078	4261	10384
0.93	43951	3701	14107	1206	27668	4671	9578
0.92	44672	2980	12359	2954	26710	5629	11563

Z výsledku je vidět správná predikce, jak se bude měnit počet chyb u obou jejich typů. U prvního typu postupně klesá, zatímco u druhého typu postupně roste. Také je vidět, že k nejméně chybám došlo při hodnotě 0.93. Kromě jednotlivých počtů chyb, je dobré se také podívat na to, u jakých jmen došlo k chybám a zda by nešlo experiment lehce upravit, aby vyšel lépe. Zejména pak se zaměřit na jména, kterých je mnohem více chybných než příjmení.

K prvním chybám druhého typu mezi jmény došlo při vzdálenost 0.98 a týkala se 4 tříd:

1. ANNAMARIA, ANNAMARINA
2. CHRISTINA, CHRISTINAM
3. CHRISTOPHER, CHRISTOPHERI
4. FRANCISCA, FRANCISCAS

Problémem 2. a 4. třídy je to, že se v nich míchají různá pohlaví. „CHRISTINA“ a „FRANCISCA“ jsou totiž ženy, zatímco „CHRISTINAM“ a „FRANCISCAS“ jsou muži. Bylo by tedy vhodné do dalšího experimentu upravit podmínky tak, aby v jedné třídě nebyly jména různého pohlaví.

Problémem 1. třídy jsou dvouslovná jména, která se někdy normují také na dvouslovná, ale někdy pouze na jednoslovná. V dalším experimentu by se tedy hodilo upravit normované tvary tak, aby se z nich bralo vždy jen první jméno. Problém 3. třídy je pak těžko řešitelný a možná zde došlo jenom k chybě při normování, jelikož rozeznat na základě změny jednoho písmena jména „KRISTIAN“ a „KRYSTOF“ je velmi složité.

Když se poté podíváme na problémy se jmény při vzdálenosti 0.97, zjišťujeme podobné problémy. Problematických tříd přibylo dalších šest:

1. ANTONIA, ANTONA, ANTONI, ANTONIE, ANTONIY, ANTONIO, ANTONII, ANTONY
2. JOHANN, JOHANA, JOHANNA
3. JOSEPH, JOSEPHA, JOSEPHI, JOSEPHY
4. MARHIAS, MARTHIAS, MARTIAS

5. MARTIN, MARTHIN, MARTINO, MARTINY, MARTINI, MARTINA, MARTINYS
6. MATHIAS, MATHAEZ, MATHAZ, MATHIAY, MATHIA, MATHIAM, ...

U 1., 2., 3. a 5. třídy je opět problém s mícháním různých pohlaví, protože máme vždy na výběr ze dvou normovaných tvarů – pro muže („ANTONIN“, „JAN“, „JOSEF“ a „MARTIN“) a pro ženy („ANTONINA“, „JANA“, „JOSEFA“ a „MARTINA“). U zbylých tříd jsou to pak problémy se zaměňováním jmen „MARTIN“ a „MATĚJ“, které však zatím dělají problém jen v jednotkách případů.

Co se týče prvních chyb, které se projevují mezi příjmeními, tak tam nejsou patrny žádné chyby, které by šli nějak jednoduše odstranit. Spíše se jedná o chyby při tvoření normovaných tvarů, protože některé normované tvary by měly patřit pod jeden sjednocený tvar. Například „MAUSIC“ a „MAUCIC“ nebo „MAKOVIC“ a „MATKOVIC“. Možná by spíše bylo při některém dalším experimentu sjednotit všechny normované tvary, které se liší o maximálně jedno písmeno.

4.1.5 Experiment č. 2

V tomto experimentu bylo zkoumáno, jak se změní počet chyb oproti experimentu č. 1 v případě, že budeme ve více jmenných normovaných tvarech brát pouze první jméno jako normovaný tvar. Například tedy jména „ANNA“ i „ANNA MARIA“ se budou nyní normovat na stejný tvar „ANNA“. Při získávání tříd se jinak používá Jaro-Winklerova vzdálenost stejně jako v prvním experimentu.

Aby tato úprava fungovala, tak stačilo provést pouze doplnění funkcí pro načítání záznamů o možnost ořezávání načítaných normovaných tvarů. Výsledky lze pak přehledně vidět v tabulce 4.2, kde jednotlivé sloupce odpovídají sloupcům v tabulce 4.1.

Tabulka 4.2: Výsledky experimentu č. 2

d_w	1. Sp.	1. Šp.	2. př. Sp.	2. př. Šp.	2. jm. Sp.	2. jm. Šp.	Celkem chyb
0.99	29769	18883	15313	0	32339	0	17883
0.98	30026	17626	15303	10	32336	3	17639
0.97	33727	13925	15255	58	32184	155	14138
0.96	38982	8670	15195	115	31696	643	9432
0.95	40905	6747	15074	239	28748	3591	10577
0.94	42011	5641	14783	530	28733	3606	9777
0.93	43903	3749	14107	1206	28332	4007	8962
0.92	44632	3020	12360	2953	27381	4958	10931

Z výsledků je patrno, že došlo ke zmenšení počtu chyb druhého typu pro jména. K jedinému většímu úbytku chyb však došlo pouze při vzdálenosti 0.96. Tento rozdíl chyb se pak projevuje i při menších vzdálenostech, ale už nijak více neroste. Vyjádřeno čísly, při vzdálenosti 0.97 se počet chyb dokonce zvětšuje, při vzdálenosti 0.96 je to už úbytek o 414 chyb a při nejmenší vzdálenosti 0.92 je to 632 chyb, což je zmenšení počtu chyb oproti vzdálenosti 0.96 pouze o 218 chyb. Prudká změna pro vzdálenost 0.96 je způsobena zejména tím, že v jedné třídě bylo v prvním experimentu 551 jmen, které měly jako normovaný tvar jméno „ANNA“, a 583 jmen, které měly naopak jako normovaný tvar jméno „ANNA MARIE“, a tyto jména už se nyní normují na stejný tvar „ANNA“.

4.1.6 Experiment č. 3

Tento experiment dále rozvíjí přechozí dva experimenty a přidává k nim kontrolu pohlaví jmen. Jeho cílem je zmenšit počet chyb druhého typu pro jména. Přidání kontroly na pohlaví byla netriviální věc. Nejdříve bylo zapotřebí doplnit funkci, která prochází všechny matriční záznamy a u nich prochází jednotlivá křestní jména, u kterých se snaží určit pohlaví. Pohlaví jména lze vyvodit u jmen jako je jméno otce, matky, ženicha nebo nevěsty, problém ale nastává u jmen narozených dětí, kde tato skutečnost nelze zjistit. Pokud je tedy použit nějaký tvar jména pouze u záznamu pro narozené dítě, je jeho pohlaví označeno jako speciální třetí pohlaví.

Kontrola na takto zjištěná pohlaví jmen je pak doplněna při porovnání dvou jmen. Ale ani tato kontrola nezabraňuje ještě vzniku tříd jmen, kde by byla jména pouze jednoho pohlaví, jelikož se jména mohou spárovat přes jména, u kterých není pohlaví identifikováno. Proto se získané třídy prochází a kontroluje se, zda v nich nejsou jména obou pohlaví. Pokud nejsou (tj. ve třídě není zároveň jméno, které je jasně určené jako mužské a které je jasně určeno jako ženské), třída se již neupravuje. Pokud je ale nalezeno jméno mužské i ženské, je rozdělena tato třída na dvě menší třídy, kde jedna obsahuje pouze mužská a druhá pouze ženská jména. Neidentifikovaná jména v této třídě se pak přidělí k jedné nebo druhé nově vzniklé třídy podle toho, k jakému tvaru jména má jméno bližší vzdálenost (tj. vypočítá se Jaro-Winklerova vzdálenost ke všem jménům a jméno se přidělí do třídy k tomu jménu, ke kterému vyšla vzdálenost nejblíže hodnotě 1).

Nad takto získanými třídami je proveden pak už stejný postup hledání chyb jako v předchozích experimentech. Výsledné hodnoty je vidět v tabulce 4.3, která je opět obdobně koncipovaná jako tabulky u předchozích experimentů.

Tabulka 4.3: Výsledky experimentu č. 3

d_w	1. Sp.	1. Šp.	2. př. Sp.	2. př. Šp.	2. jm. Sp.	2. jm. Šp.	Celkem chyb
0.99	29769	17833	15313	0	32339	0	17833
0.98	30026	17626	15303	10	32337	2	17638
0.97	33726	13926	15255	58	32335	4	13988
0.96	38091	9561	15195	118	32332	7	9686
0.95	40901	6751	15074	239	32326	13	7003
0.94	41963	5689	14783	530	31529	810	7029
0.93	43855	3797	14107	1206	31139	1200	6203
0.92	44559	3093	12360	2953	30202	2137	8183

Z výsledných hodnot jde vidět velké zlepšení. Počet chyb druhého typu pro jména je při vzdálenosti 0.95 pouhých 13, což je v rámci celkového počtu 32339 jmen, pro které existuje normovaný tvar, zanedbatelný počet (méně než 0.1 %). K první větší chybě dochází při vzdálenosti 0.94, kdy dochází k 810 chybám. Tento problém je způsoben blízkostí jmen, která se normují na normované tvary „MATEJ“ (2596 jmen) a „MARTIN“ (745 jmen). K podobné chybě dochází dále i při vzdálenosti 0.93. Zde je celkový počet chyb 1200 a kromě výše zmíněného je zde největším problémem blízkost jmen, která se normují na tvary „JIRI“ (reálná jména jsou různé variace jména „GEORGE“ a je jich celkem 1600) a „REHOR“ (variace jména „GREGOR“ s celkovým počtem 389 jmen). U posledně měřené vzdálenosti 0.92 je pak celkem již 2137 chyb. Opět je to ale způsobené dalším dvojicí podobných jmen, tentokrát je to dvojice normovaných tvarů „MARIE“ (3958 jmen) a „MARKETA“ (917 jmen).

Nebýt výše zmíněných třech trojic jmen, tak by celkový počet chyb druhého typu pro jména byl pro vzdálenost 0.92 pouhých 86 jmen, což je v porovnání s celkovým počtem jmen pořád necelá třetina procenta chybných jmen.

Co se týče celkového počtu chyb, tak pro vzdálenost 0.93 dostáváme nejmenší počet chyb 6203, což je vůči celkovému počtu 47652 jmen a příjmení asi 13 % chybných přiřazení, přičemž víme, že asi 2.5 % z toho jsou zaviněna dvojicemi „MATEJ“,„MARTIN“ a „JIRI“, „REHOR“.

4.1.7 Experiment č. 4

Pro porovnání s předchozími experimenty byl ještě proveden pokus vyzkoušet při počátečním třídění do tříd jinou vzdálenost než Jaro-Winklerovu vzdálenost a to konkrétně Hammingovu a Levenshteinovu vzdálenost. Původně bylo v plánu vyzkoušet je pro různé maximální vzdálenosti, ale ukázalo se, že už pro maximální vzdálenost 2 nelze vytvořit třídy jmen popsaným způsobem výše, jelikož dojde k vytvoření jen několika málo tříd o velkých počtech jmen.

Jediné rozumné výsledky tak šlo dostat pouze pro maximální vzdálenost 1. Pro Hammingovu vzdálenost tak bylo objeveno 11322 chyb prvního typu, 756 chyb druhého typu pro příjmení a 83 chyb druhého typu pro jména. Z velkého počtu chyb prvního typu je vidět, že bylo vytvořeno méně tříd a pro jeden normovaný tvar tak existuje více tříd. Celkový počet chyb je 12161, což je zhruba dvakrát tolik chyb než pro Jaro-Winklerovu vzdálenosti 0.93 z předchozího experimentu.

Pro Levenshteinovu vzdálenost bylo nalezeno 7980 chyb prvního typu, 1773 chyb prvního typu pro příjmení a 2633 chyb druhého typu pro jména. Celkový počet chyb je pak 12386. Z toho je vidět, že počty chyb u jednotlivých typů se pro Levenshteinovu a Hammingovu vzdálenost liší, ale celkový počet je podobný a nedostačující výsledkům, které byly získány pro Jaro-Winklerovu vzdálenost.

4.2 Spojování záznamů

Tato část se zabývá již přímým problémem spojení dvou genealogických záznamů. Budou se v ní provádět experimenty na datech za účelem nalezení, co největšího počtu správných párování, přičemž se bude využívat získaných zkušeností z předchozích experimentů.

Hlavním cílem v každém experimentu bude spojit údaje o základní rodině. Základní rodinnou se rozumí rodiče a jejich děti. Z matriky narozených a oddaných bude tedy zapotřebí najít a spojit tyto tři typy dvojic záznamů:

- záznam o narození otce a záznam o jeho oddání
- záznam o narození matky a záznam o jejím oddání
- záznam o narození dítěte a záznam o oddání jeho rodičů

Jako základní záznam, ke kterému se budou dohledávat ostatní, se bude tedy vždy brát jeden záznam o oddání. Dále bude tedy rozebráno, jak se budou jednotlivé údaje v tomto záznamu porovnávat s údaji v záznamech o narození.

4.2.1 Párování se záznamy o narození rodičů

Při spojování záznamu o oddání a narození rodiče si je potřeba uvědomit, že vždy se tímto způsobem má k záznamu o oddání napojit maximálně jeden záznam o narození a to ten nejpravděpodobnější záznam, jelikož se každý z rodičů mohl narodit pouze jednou. To je jeden z rozdílů oproti spojování záznamu o oddání se záznamy o narození dětí, jelikož dětí je většinou v jednom manželství více. Záznamy, pro které je pravděpodobnost menší, ale dostatečně vysoká, si ale může program přesto zapamatovat jako záložní záznamy pro případ, že uživatel později rozhodne, že primární spojení bylo vytvořeno nesprávně.

Když se tedy srovnávají dvojice záznamů o narození rodiče a o jeho oddání, je vždy možné porovnávat následující společné údaje:

- údaj o jméně a příjmení u záznamu o narození s údajem o jméně a příjmení ženicha nebo nevěsty
- údaj o jméně otce u záznamu o narození s údajem o jméně otce ženicha nebo nevěsty
- údaj o jméně a příjmení matky u záznamu o narození s údajem o jméně a příjmení matky ženicha nebo nevěsty
- údaj o datu narození s údajem o datu oddání
- údaj o zaměstnání otce u záznamu o narození s údajem o zaměstnání otce ženicha nebo nevěsty
- údaj o bydlišti u záznamu o narození s údajem o bydlišti ženicha nebo nevěsty

První tři typy dvojic údajů se mohou porovnávat například za pomoci tříd jmen, které byly tvořeny v experimentech v předchozí sekci. Tyto dvojice údajů také budou tvořit základ porovnání v následujících experimentech.

Čtvrtý typ dvojice údajů bude také důležitou součástí následujících experimentů, jelikož bude často rozhodovat o spojení v případě, že bude existovat více možností. U tohoto typu si bude potřebovat na začátku správně nadefinovat, pro jaké rozdíly v datech ještě uvažovat možné spojení a pro jaké už naopak ne. Navíc u novějších záznamů lze využít navíc i údajů o věku ženicha a nevěsty a dle toho upravit nadefinované preference.

V následujících experimentech budou tedy platit následující pravidla pro porovnání tohoto typu údaje pro ženicha:

- datum narození ženicha může být maximálně 60 let před datem oddání
- datum narození ženicha musí být minimálně 15 let před datem oddání
- pro všechny data narození v tomto období jsou přiřazeny různé váhy s tím, že největší váha je kladena tak, aby věk ženicha v době oddání byl mezi 20 a 25 lety
- pokud je zadaný i věk ženicha, musí pak platit, že po odečtení tohoto věku od roku oddání se může rok narození ženicha lišit maximálně o 2 roky od takto zjištěného roku

Pro nevěstu pak budou platit podobná pravidla:

- datum narození nevěsty může být maximálně 40 let před datem oddání (starší nevěsty nás v těchto experimentech příliš nezajímají, protože už nemůžou mít děti, a nemohou tak na sebe vázat další záznamy)

- datum narození nevěsty musí být minimálně 15 let před datem oddání
- pro všechny data narození v tomto období jsou opět přiřazeny různé váhy s tím, že největší váha je kladena tak, aby věk nevěsty v době oddání byl mezi 18 a 24 lety
- pokud je zadaný i věk nevěsty, musí pak platit, že po odečtení tohoto věku od roku oddání se může rok narození nevěsty lišit maximálně o 2 roky od takto zjištěného roku

Zbylé dva typy dvojic údajů už se nebudou tolik využívat, protože nemusí být podle zkušeností shodné. V následujících experimentech se tedy bude spíše počítat, pro kolik spárovaných dvojic záznamů si i tyto údaje odpovídají a pro kolik naopak ne.

4.2.2 Párování záznamů o narození dětí

Jak už bylo zmíněno výše, tak dětí může být v manželství více a při spárování se tedy budou spojovat všechny záznamy. Při srovnávání záznamu o oddání rodičů se záznamem o narození dítěte se pak mohou porovnávat následující údaje:

- údaj o příjmení dítěte s údajem o příjmení ženicha
- údaj o jméně otce dítěte s údajem o jméně ženicha
- údaj o jméně a příjmení matky dítěte s údajem o jméně a příjmení nevěsty
- údaj o datu narození dítěte s údajem o datu oddání
- údaj o zaměstnání otce dítěte s údajem o zaměstnání ženicha
- údaj o bydlišti u dítěte s údajem o bydlišti ženicha

Podobně jako u párování záznamů o narození rodičů i zde se budou porovnávat v experimentech zejména první tři typy dvojic údajů a bude se k tomu využívat zejména tříd jmen vytvářených v předchozích experimentech.

Stejně tak se bude využívat i čtvrtého typu dvojic údajů, kdy si bude potřeba správně nadefinovat, kolik let po oddání se mohou rodit děti. A i zde se bude u pozdějších údajů využívat údaje o věku nevěsty.

V následujících experimentech budou tedy platit následující pravidla pro porovnání tohoto typu údaje:

- datum narození dítěte může být maximálně 30 let po svatbě (nevěstě je mezi 15 a 45 lety)
- k narození dítěte může dojít nejdříve v roce oddání
- pro všechny data narození v tomto období jsou přiřazeny různé váhy s tím, že největší váha je hned v roce oddání a postupně se každým rokem snižuje
- pokud je zadaný i věk nevěsty, zmenšuje se maximální hranice na 45 mínus věk nevěsty let (např. pokud je nevěstě 28 let, může mít dítě maximálně 17 let po oddání)

Zbylé dva typy dvojic údajů se budou opět využívat v následujících experimentech pro zjišťování, pro kolik spárovaných dvojic záznamů si tyto údaje odpovídají a pro kolik naopak ne.

4.2.3 Problematika vdov

Předtím než se přejde k experimentům je ještě potřeba si probrat problematiku záznamů, kde nevěstou je vdovou. Tyto záznamy totiž obsahují trochu jiné údaje než ostatní záznamy o oddání. Platí pro ně, že nemají údaj o příjmení nevěsty a ani žádný údaj o jejích rodičích. Obsahují však navíc údaje o jméně a příjmení jejich předchozího manžela.

Před spojováním záznamů o oddání se záznamy o narození je tedy ještě vhodné zkusit najít pro vdovy záznam o jejich prvním manželství a doplnit z něj informace o rodičích a jejím příjmení k jejímu druhému manželství.

Při tomto spojování záznamu o dalším manželství se záznamem o prvním manželství nevěsty se porovnávají následující čtyři údaje:

- údaje o jméně nevěsty u obou záznamů o oddání
- údaj o jméně vdovce s údajem o jméně ženicha z prvního manželství
- údaj o příjmení vdovce s údajem o jméně ženicha z prvního manželství
- údaje o datu oddání u obou manželství

Aby dva záznamy byly prohlášený za spojené, tak musí dojít k tomu, že se první tři typy údajů neliší a maximálně nejde porovnat jenom jedna dvojice, kde nějaký údaj chybí (nejčastěji chybí údaj o příjmení vdovce). Porovnávání jmen bude v experimentech probíhat opět za pomoci tříd vytvořených v předchozích experimentech jako i při spojování jiných záznamů. Údaje o datu oddání se pak mohou lišit maximálně o 40 let (datum oddání pro druhé manželství nemůže samozřejmě předcházet datu oddání druhého manželství).

4.2.4 Experiment č. 5

V tomto pokusu se pracuje pouze s normovanými tvary jmen, které jsou k dispozici ve zdrojových datech. Cílem tohoto pokusu pak je identifikovat vztahy v základní rodině. Tyto zjištěné vztahy si poté u jednotlivých záznamů o narození uložit a použít při dalších experimentech pro porovnání a zjištění počtu pravděpodobně špatně spárovaných záznamů.

Dva záznamy jsou v tomto experimentu prohlášeny za sobě odpovídající jen v případě, že se shodují všechny odpovídající si normované tvary jmen. V případě většího počtu možností u matky a otce je vybrán ten záznam, který odpovídá nejlépe dle dat záznamů, přičemž se vychází z pravidel uvedených v předchozích dvou podsekcích 4.2.1 a 4.2.2. Jediná úprava normovaných tvarů, která se provádí, je ta, že u víceslovných normovaných tvarů se zahazuje vše kromě prvního slova, tj. „ANNA“ a „ANNA MARIA“ se berou jako dvě shodná jména. Výsledky tohoto pokusu pak lze přehledně pozorovat v tabulce 4.4.

Tabulka 4.4: Výsledky experimentu č. 5

	nalezení	nenalezení
otcové	733	466
matky	748	451
děti	4275	4252

Při pohledu na výsledky a relativně nižší počet úspěšných spojení záznamu o narození rodiče se záznamem o jeho oddání je potřeba si uvědomit následující souvislosti. Pro brzká

oddání před rokem 1700 nejsou k dispozici žádné záznamy o narození starší 15 let. V tomto případě jde o 108 záznamů, ke kterým nelze nikdy nic spárovat. Dále pak pro záznamy do roku 1710 (dalších 79 záznamů) je také ještě malá pravděpodobnost na dohledání. Na zdrojových datech se to povedlo pouze pro 8 ženichů a 13 nevěst (to také odpovídá tomu, že se nevěsty vdávají v mladším věku. Po odečtení těchto brzkých záznamů bylo tedy úspěšně spojeno zhruba 65 % záznamů o narození ženichů a zhruba 66 % záznamů o narození nevěst, tedy zhruba dvoutřetinová úspěšnost. Pro zbylou třetinu záznamů může být důvodem nenalezení záznamu to, že se daný člověk přistěhoval odjinud, jeho příjmení u narození nemělo zadaný normovaný tvar nebo bylo toto normování provedeno nepřesně.

Při spojování záznamů o narození dětí se záznamy o oddání rodičů dostáváme úspěšné spojení pro zhruba polovinu dětí z těch, které mají ve zdrojových datech normovaný tvar pro příjmení (8527 dětí). Zde je také podobný problém u brzkých záznamů o narození, jelikož svatba rodičů mohla proběhnout před rokem 1685, od kterého jsou až k dispozici zdrojová data o narození. Dalšími problémy pak opět může být to, že se rodiče oddali někde jinde, či opět to může být nepřesnostmi při normování.

Důležité pak je ještě zmínit výsledky spojování záznamů o oddání pro vdovy. Zde se podařilo dohledat předchozí manželství pro 278 vdov z celkového počtu 369 procházených, takže pro zhruba 75 % vdov.

Jak je zmíněno výše, tak v tomto pokusu bylo také měřeno, u kolika úspěšných spojení se shodovalo číslo domu bydliště a u kolika alespoň vesnice. Aby si název vesnice odpovídal, tak musel být název pro oba dva záznamy úplně stejný, tedy zde nebyly brány do úvahy různé tvary jako u jmen a příjmení. Tato shoda pak byla měřena mezi těmito dvojicemi záznamů:

- mezi záznamem o narození otce a záznamem o narození jeho dítěte
- mezi záznamem o narození matky a záznamem o narození jejího dítěte
- mezi jednotlivými záznamy o narození dětí (tj. mezi sourozenci), kde za správné číslo bydliště se bralo to, které bylo u největšího množství dětí a tento počet byl větší než 1
- mezi záznamem o oddání (brala se adresa otce) a jeho záznamem o narození
- mezi záznamem o oddání (brala se adresa matky) a jejím záznamem o narození
- mezi záznamem o oddání (brala se adresa otce) a záznamy o narození jeho dětí

U prvních dvou typů dvojic byla shoda v čísle domu minimální, konkrétně pro otce byla shoda s adresou dítěte nalezena pouze ve 13 případech, u matek pak byla shoda v 10 případech, naopak odlišné číslo domu bylo v 1146 případech (resp. 1281), ve zbylých 1637 (resp. 1690) případech pak alespoň u jednoho záznamu chyběl údaj o čísle domu. Shoda pro číslo domu mezi jednotlivými záznamy o narození dětí byla už o něco větší, ale pořád k ní došlo u pouhých 149 záznamů, u dalších 80 záznamů to pak nešlo rozeznat, protože šlo o jedináčky, u 2189 záznamů ke shodě nedošlo a 1957 nebylo uvedeno číslo domu.

U zbylých třech typů dvojic, kdy jedním z dvojice byl vždy záznam o oddání, byly výsledky pro číslo domu ještě horší. Pro otce došlo ke shodě pouze ve třech případech, pro matky ve čtyřech případech a pro děti pak ve 26. To potvrzuje domněnku, že se nelze spolehnout na spojování záznamů dle tohoto údaje.

Měření shody pro názvy vesnic už byly mnohem větší. U prvních dvou typů dvojic záznamů byla shoda nalezena u 2054 případů (resp. 1864). Rozdílných případů pak bylo 661 (resp. 1053), což činí úspěšnost shody asi 76 % (resp. 64 %). Větší shoda u mužů může odpovídat tomu, že se spíše ženy stěhovaly za svým mužem, pokud byl z jiné vesnice. Mezi jednotlivými záznamy o narození dětí došlo ke shodě vesnice v 3871 a pouze v 381 případech došlo k rozdílu. To znamená, že ke shodě došlo ve zhruba 91 %.

U zbylých třech případů byly výsledky podobné, akorát pro množství záznamů o oddání není dostupná informace o vesnici, takže bylo i velké množství dvojic záznamů, pro které nešlo určit shodu. Pro otce tedy došlo celkem ke 278 shodám a 90 rozdílům, tj. byla shoda u zhruba 76 % dvojic, pro matky byla čísla podobná: 278 shod, 86 rozdílů (76 %). Pro děti pak byla shoda v 1626 případech, zatímco rozdíl byl u 800 případů (67 %). Tento údaj se zde všech měřených tedy jeví nejlépe a mohl by se zapojit do rozhodování o spojení dvou záznamů například v případě, kdy nějaké dítě má podobnou pravděpodobnost, že patří do dvou různých rodin, kde každá žije v jiné vesnici.

Posledním údajem měřeným u tohoto spojení byl počet shodných povolání. Podobně jako u názvu vesnice i v tomto případě musely být názvy povolání naprosto stejné. Povolání se pak porovnávaly na rozdíl od předchozích údajů pouze pro tři různé dvojice záznamů:

- mezi záznamem o oddání (bralo se povolání otce ženicha) a ženichovým záznamem o narození
- mezi záznamem o oddání (bralo se povolání otce nevěsty) a nevěstiným záznamem o narození
- mezi záznamem o oddání (bralo se povolání ženicha) a záznamy o narození dětí

Výsledky měření pak ukazují, že počet shod je větší než, jaký byl pro měření shod u čísel domů, ale není zase tak vysoký jako při měření shod názvů vesnic. Konkrétně pro první typ dvojic záznamů bylo nalezeno 71 shodných případů a 246 rozdílných případů tzn., že úspěšnost je pouhých zhruba 22 %. Pro druhý typ dvojic záznamů je počet shodných případů 121, zatímco rozdílných případů je 203 (tj. úspěšnost asi 37 %). Pro poslední typ je pak počet shodných případů 439 a rozdílných 235, což činí nejlepší úspěšnost asi 65 %. Z těchto výsledků je vidět, že ani tento údaj se nedá při spojování záznamů použít.

4.2.5 Experiment č. 6

Tento experiment vychází z výsledků experimentů č. 3. Při spojování záznamů se tedy využívá tříd jmen vzniklých v tomto experimentu. Jeho cílem je pak ukázat, jaké výsledky dostaneme, pokud se při spojování dvou záznamů budou brát za shodné pouze ty řetězce, které spadají do stejné třídy jmen. Zkoumat výsledky se budou pro porovnání pro všechny vzdálenosti, pro které byly v experimentu č. 3 tyto třídy vytvářeny.

Aby byly v tomto experimentu dva záznamy spojeny, tak musí platit, že je nalezena shoda u všech dvojic porovnávaných jmen. V případě, že existuje více takových možností pro narození rodiče, tak se jako nalezený rodič označí ten, pro který odpovídá lépe věkový rozdíl, přičemž se vychází z pravidel v podsekcích 4.2.1 a 4.2.2.

Výsledky tohoto experimentu pak lze vidět přehledně v tabulce 4.5, kde ve 2., 4. a 5. sloupci je vždy vidět počet stejně spárovaných záznamů jako pro normované tvary (Ok) a počet odlišně spárovaných záznamů (Ko).

Tabulka 4.5: Výsledky experimentu č. 6

d_w	Děti Sp./Šp.	Bezdětné	Otcové Sp./Šp.	Matky Sp./Šp.
0.99	1460/2816	783	83/553	89/628
0.98	1599/2678	751	86/551	108/610
0.97	2075/2209	645	158/484	197/528
0.96	2625/1671	527	235/409	317/417
0.95	2960/1355	463	318/327	400/337
0.94	3228/1116	428	377/285	467/274
0.93	3504/895	384	409/264	536/513
0.92	3609/919	350	418/278	532/255

Z výsledků lze, že nejlepších hodnot se dosahuje pro vzdálenost 0.93. Při této vzdálenosti je zhruba 82 % záznamů o narození dětí, které se spárují ke stejným manželstvím jako pro normované tvary jmen. Podobně je zhruba 56 % záznamů o narození ženichů stejně spárováno a také je zhruba 72 % záznamů o narození nevěst spárováno stejně.

4.2.6 Experiment č. 7

V tomto experimentu se používají podobně jako u předchozího experimentu třídy vytvářené v experimentu č. 3. Tentokrát se však vždy bude využívat tříd vytvořených pro Jaro-Winklerovu vzdálenost 0.93. Rozšířením pak bude možnost navázání spojení dvou záznamů i v případě, že některá dvojice odpovídajících si jmen u dvou porovnávaných záznamů nebude ležet ve stejné třídě. Toto však může nastat pro maximálně jednu dvojici jmen a musí pro ni platit stále alespoň nějaká minimální zvolená hraniční vzdálenost (např. 0.8).

V tabulce 4.6 pak lze vidět vývoj počtu stejně a různě spárovaných dětí, ženichů a nevěst při porovnání s párováním pomocí normovaných tvarů.

Tabulka 4.6: Výsledky experimentu č. 7a pro třídy s $d_w = 0.93$

d_w	Děti Sp./Šp.	Bezdětné	Otcové Sp./Šp.	Matky Sp./Šp.
0.92	3558/841	376	412/261	542/207
0.90	3588/811	372	422/252	554/195
0.88	3637/780	367	439/238	566/187
0.86	3690/735	357	460/222	576/182
0.84	3728/710	351	479/204	589/172
0.82	3750/691	351	493/201	602/164
0.80	3753/733	344	497/203	600/179

Z výsledků je patrno významné zlepšení. Nejlepší hodnoty vycházejí pro minimální vzdálenosti 0.81 a 0.82. Pro párování záznamů o narození dětí se záznamy o oddání se při těchto vzdálenostech zvedl počet na přibližně 88 % stejně spárovaných záznamů jako pro normované tvary jmen, což je zvýšení zhruba o 6 %. Ještě k většímu zlepšení však došlo při párování záznamů o narození ženichů se záznamy o oddání. Zde došlo ke zvětšení o asi 11 % procent na 67 % stejně spárovaných záznamů. Pro nevěsty pak došlo ke zlepšení o 8 % na 80 % stejně spárovaných záznamů.

Pro srovnání byl tento experiment učiněn ještě pro třídy, které byly vytvořeny pro Jaro-Winklerovu vzdálenost 0.94. Výsledky pro vybrané minimální vzdálenosti je pak vidět

v tabulce 4.7. Nejlepších výsledků je pak dosáhnuto pro minimální vzdálenost 0.78, ale z hlediska procentuální úspěšnosti shodných spárování jsou pro děti jen zhruba na úrovni výsledků z předchozího experimentu pro výchozí vzdálenost 0.93. Nejlepší výsledky jsou pro otce, kdy jsou na úrovni vzdálenosti 0.82 z předchozího experimentu, takže ani ty nejsou lepší.

Tabulka 4.7: Výsledky experimentu č. 7b pro třídy s $d_w = 0.94$

d_w	Děti Sp./Šp.	Bezdětné	Otcové Sp./Šp.	Matky Sp./Šp.
0.82	3573/838	371	480/200	568/188
0.80	3580/875	364	486/208	575/193
0.78	3588/876	361	492/208	582/196
0.75	3594/897	358	501/209	584/209

4.2.7 Další experimenty a měření

V rámci předchozího pokusu byly ještě měřeny dvě zajímavější statistiky. První z nich byla statistika počtu nalezených generací od daného záznamu o oddání. Pro tento záznam o oddání pak muselo platit, že pro rodiče ženicha a nevěsty už nešel dohledat záznam o oddání. Tato statistika byla spočítána, jak pro normované, tak pro nenormované tvary jmen.

Pro normované tvary se podařilo najít nejvíce 5 na sebe navazujících záznamů o oddání a to pro celkem 11 počátečních záznamů o oddání. Podařilo se tak například propojit záznamy pro rod „HUBENI“, které byly od sebe vzdálené 159 let, kde výchozím záznamem byl sňatek již z roku 1689 (tedy jeden z nejstarších zdrojových záznamů), kdy si Martinus Hubeni vzal Helenu Krisanich a poslední záznamem byl sňatek z roku 1848 (tedy naopak jeden z nejnovějších záznamů), kdy si jejich prapravnučka Anna Hubeni vzala Franze Slunski.

Pro nenormované tvary se dokonce zdánlivě povedlo najít i 6 po sobě jdoucích generací, ale při bližším průzkumu bylo zjištěno, že jedna dvojice záznamů byla špatně spárována. Největším rozpětím tak bylo 5 generací a to pro opět pro stejný případ rodiny Hubeni jako v rámci normovaných tvarů.

Přehledně lze všechny jednotlivé počty generací pro normované i nenormované tvary vidět v tabulce 4.8.

Tabulka 4.8: Výsledky experimentu s generacemi

počet generací	normované	nenormované
1	444	507
2	91	104
3	38	32
4	17	17
5	11	5
6	0	1

Druhá statistika byla zaměřena na zpřesnění rozdělení vah rozdílu dat dvou spárovaných záznamů, tzn., že jejím cílem bylo zjistit, kolik let po svatbě se narodilo kolik dětí, v kolika se vdávaly ženy a v kolika se ženili muži. Tato statistika byla zkreslena již prvotně navrženým

rozvržením vah, pokud by však byl počáteční odhad významněji odlišný od skutečnosti, ve statistice by se to projevilo a při následném slučování záznamů by se výrazněji navýšil počet úspěšných spojení.

Výsledky tohoto měření ale ukázaly, že prvotní odhad nebyl natolik mylný, aby bylo potřeba váhy příliš měnit. Přesto byla na jejich základě poté provedena úprava vah pro jednotlivé rozdíly dat a bylo znovu vypočítáno, kolik bylo tentokrát nalezeno správných spojení. Výsledné hodnoty však byly lepší jen o zanedbatelný počet jednotek správných dvojic.

4.3 Celkové zhodnocení

V prvních čtyřech experimentech se podařilo zjistit problémy při tvorbě tříd odpovídajících si jmen, při kterých se podařilo snížit počet chyb z 9578 na 6203, tedy o více než třetinu. Důležitou roli v tomto snížení hrála především informace o pohlaví jmen.

Dále pak bylo z podrobného výpisu chyb zjištěno, že pouhou úpravou dvou tříd (jejich rozdělením na dvě menší třídy) lze tento počet snížit ještě o dalších asi 1500 chyb. To při celkovém počtu 47652 jmen dává úspěšnost přes 90 %.

Jako nejlepší metrika se pro toto vytváření tříd osvědčila Jaro-Winklerova vzdálenost. Pro tuto metodu lze totiž mnohem lépe vytvářet menší kompaktnější shluky jmen než například pro Levenshteinovu metodu, kde už pro vzdálenost 2 se vytvářejí velmi rozsáhlé shluky.

Ve zbylých experimentech se poté podařilo s těmito třídami pracovat, provést úspěšně spárování mnoha záznamů a naměřit z nich další údaje. Konkrétně se podařilo spárovat 88 % záznamů o narozených dětech v daném manželství stejně jako pro normované tvary jmen, 67 % záznamů o narození ženichů a 80 % o narození nevěst.

Dále se zjistilo, že údaje o čísle domu v adresách nejsou příliš věrohodné a nelze je použít při párování. Podobně pak nevyhovují údaje o práci daných lidí, pro tyto údaje by možná pomohlo, kdyby se vytvořil nějaký pomocný převod na jednotný tvar, protože některé názvy povolání měli více možných úplně syntakticky rozdílných názvů. Jediným údajem, který by šel použít při spojování, je tak údaj o vesnici. Ten se nejvíce osvědčuje při porovnávání dětí mezi sebou, kdy pro více než 90 % dětí, že se všechny v dané rodině narodily ve stejné vesnici.

4.4 Návrhy do budoucna

V experimentech pro tvorbu tříd jmen byly ukázány případy dvojic jmen jako Martin a Matěj, Řehoř a Jiří či Marie a Markéta, pro které platilo, že shluky tvarů těchto jmen měli k sobě blízkou vzdálenost, a proto byly sloučeny do jedné společné třídy jmen. Při dalším rozšiřování modulu na tvorbu tříd by se tedy mělo hledět na možnost nastavit při počátečním vstupu, které tvary jmen si nesmí odpovídat. Tím pádem by se tyto tři dvojice jmen nesloučily do jedné třídy, ale existovaly by dvě samostatné, pro které už by byl počet chyb menší.

Opačným případem je naopak problém rozpadu jednoho příjmení, které v průběhu let mění úplně svůj syntaktický tvar, do dvou oddělených tříd pro příjmení. To se děje často například z jazykového důvodu, kdy německé jméno Schon se v novější záznamech objevuje již pod českým jménem Pěkný. Na vstupu by si tedy měl mít možnost uživatel nastavit, které dvojice jmen se mají brát za stejné a podle toho pak sjednotit vytvořené třídy.

Další možností na rozšíření je možnost vytvoření nějakého nového speciálního fonetického algoritmu. V současném programu dochází při načtení dat akorát k nahrazení písmene W za V a písmene Y za I, což je pouhý základ. Spoustu příjmení však končí například na různé podobné tvary TZ, TS, TSH, TSCH, které by mohly být nahrazeny jedním společným tvarem. Dále se občas objevují i zdvojená písmena, ty by se pak mohli nahradit za pouhé jedno dané písmeno.

Pokud by se v budoucnu podařilo získat větší množství zdrojových dat z matriky zemřelých, lze realizovat rozšíření současné podpory pro tento typ záznamů. Po doplnění metody na porovnávání záznamu o narození se záznamem o úmrtí by pak šlo zlepšit i metodu na porovnávání záznamu o narození se záznamem o oddání. V ní by se doplnila kontrola na to, že ženich a nevěsta nesmí být po smrti, což by mohlo výrazně vylepšit výsledky.

Kapitola 5

Závěr

Cílem této práce bylo prostudovat metody slučování dat v matričních dokumentech a pro dodaná data pak identifikovat problémy při identifikaci rolí osob v nich obsažených.

Proto byla v této práci nejdříve shrnuta obecná teorie, která slouží ke spojování jakýchkoliv záznamů z více souborů dat. Byly podrobně rozebrány jednotlivé metody na porovnávání řetězců a to jak metody, které vycházejí z úprav jednoho řetězce na jiný, tak metody založené na fonetických shodách. Také byl probrán vliv historických událostí a rozdělení pravděpodobnosti spojité náhodné veličiny.

V následující části pak byl stručně popsán návrh a implementace jednotlivých základních modulů programu, který byl využit pro experimentování a některé jeho metody mohou být dále využity i jako základ pro finální program na slučování genealogických dat.

V poslední obsáhlejší části pak byly popsány jednotlivé experimenty, které byly prováděny nad vstupními daty, a také bylo provedeno vyhodnocení jejich výsledků. Tyto výsledky jsou důležitým podkladem pro další pokračování v této oblasti.

Při dalším pokračování v této oblasti by bylo vhodné více prověřit stávající fonetických algoritmů a navrhnout případně podobný vlastní fonetický algoritmus, který bude odpovídat jihoslovanské skupině slovanských jazyků, což je jazyková skupina, ze které vycházejí jména a příjmení z dodané sady dat. Pro další rozvoj v této oblasti by bylo také vhodné získat více vstupních data z matriky zemřelých. Tyto data totiž nebyla pro tuto prací v současnosti úplná, a nešlo tak naplno využívat například faktoru vysoké kojenecké úmrtnosti, který by dokázal vyloučit podstatnou část záznamů o narození z dalšího porovnávání.

Výsledně vytvořený program pro slučování dat by pak měl po dalším pokračování zvládat i různá vlastní nastavení od uživatele, například možnost vložení vlastního fonetického algoritmu pro jinou jazykovou skupinu.

Literatura

[1] *Obyvatelstvo - roční časové řady*. [Online; navštíveno 01.05.2019].
URL https://www.czso.cz/csu/czso/obyvatelstvo_hu

[2] *Historická demografie*. 11, Ústav československých a světových dějin ČSAV, 1987.

[3] Fajmon, B.; Hlavičková, I.; Novák, M.; aj.: *Numerická matematika a pravděpodobnost*. VUT v Brně, 2014.

[4] Fellegi, I. P.; Sunter, A. B.: *A theory of record linkage*. 1969.

[5] Fialová, L.; aj.: *Dějiny obyvatelstva českých zemí*. Mladá fronta, 1998, ISBN 80-204-0720-0.

[6] Hordějčuk, V.: *Algoritmus Wagner-Fischer*. [Online; navštíveno 01.05.2019].
URL http://voho.eu/wiki/algoritmus-wagner-fischer/?fbclid=IwAR1ep2xBwMYEyu0be0bjQWFk3aEqyVSTriuF98iuuDJG_cHV804dk7VuFTk

[7] Pejčoch, D.: *Využití Fuzzy Match algoritmu pro čištění dat*. [Online; navštíveno 01.05.2019].
URL https://docplayer.cz/4472255-Vyuziti-fuzzy-match-algoritmu-pro-cisteni-dat.html?fbclid=IwAR3RDNuxH5dPAidSTmsuMweDrXhpY6XTSaIjel1oMaV6V8UnQ2jZaQ7Vhxw

[8] Smetanin, N.: *Phonetic algorithms*. [Online; navštíveno 01.05.2019].
URL
https://ntz-develop.blogspot.com/2011/03/phonetic-algorithms.html?fbclid=IwAR1U8OuLqzhqcO_SJQ0Tcxn4kh_jc7ERw2S7epH6Dl3392MnivToif2KLy0

Příloha A

Obsah CD

- `/src/` – složka se zdrojovými soubory
- `/data/` – složka se zdrojovými daty
- `/doc/` – složka se zdrojovými soubory písemné práce
- `/README.txt` – pokyny k překladu a spuštění experimentů
- `/xsormj00-BP.pdf` – elektronická verze písemné zprávy
- `/plakat.pdf` – plakát shrnující práci

Příloha B

Zdrojová data

V této příloze jsou detailně popsána data, na kterých bylo testováno spojování genealogických záznamů. Data původem pocházejí z matrik obce Frelichov (nyní Jevišovka) z časového rozmezí 1686 až 1850. Z těchto matrik byla přepsána vedoucím práce doc. Ing. Františkem Zbořilem, Ph.D. do formátu CSV (Comma-separated values) a poskytnuta pro experimenty v této práce.

V následujících sekcích jsou rozebrány jednotlivé druhy záznamů podle toho, z jakého druhu matriky pocházejí. U každého záznamu jsou popsány pouze údaje, které byly v experimentech využívány, nebudou se zde tedy rozebírat například sloupce se s údaji o svědcích na svatbě, které nebyly použity. Pro každý sloupec o jméně pak platí, že k němu existuje sloupec s normovaným tvarem tohoto jména.

Matriky narozených

Ideální záznam z matriky narozených poskytuje tyto údaje:

- **Datum narození** – je rozdělené do tří samostatných polí, které obsahují den, měsíc a rok uvedený vždy v číselné podobě.
- **Obec a číslo domu** – jsou rozdělené do dvou samostatných polí. Obec bývá uvedena zkratkou a je u valné většiny záznamů a číslo domu je v číselné podobě a uvedené ve zhruba polovině případů.
- **Jméno a příjmení dítěte** – jsou rozdělené do dvou samostatných polí uvedené vždy v řetězcové podobě.
- **Jméno a povolání otce** – jsou rozdělené do dvou samostatných polí uvedené vždy v řetězcové podobě. Povolání otce však zhruba v polovině případů chybí.
- **Jméno a rodné příjmení matky** – jsou rozdělené do dvou samostatných polí uvedené vždy v řetězcové podobě. Rodné příjmení matky však zhruba v polovině případů chybí.
- **Jméno a povolání otce matky** – jsou rozdělené do dvou samostatných polí uvedené vždy v řetězcové podobě. Oba údaje však chybí ve většině záznamů.

Matriky zemřelých

Ideální záznam z matriky zemřelých poskytuje tyto údaje:

- **Datum úmrtí** – je rozdělené do tří samostatných polí, které obsahují den, měsíc a rok uvedený vždy v číselné podobě.
- **Typ** – je jedno samostatné řetězcové pole. Obsahuje informaci o tom, zda se jednalo o dítě, manželku, apod. Tato informace bývá ukryta pod různými zkratkami, takže je potřeba nejdříve zjistit, co která značka znamená. Chybí zhruba ve čtvrtině případů.
- **Jméno a příjmení** – jsou rozdělené do dvou samostatných polí uvedené vždy v řetězcové podobě.
- **Jméno a povolání otce** – jsou rozdělené do dvou samostatných polí uvedené vždy v řetězcové podobě. Jméno se nachází u většiny záznamů, ale povolání otce chybí zhruba v polovině případů.
- **Obec a číslo domu** – jsou rozdělené do dvou samostatných polí. Obec bývá uvedena zkratkou a je uvedena u většiny záznamů. Číslo domu je v číselné podobě a uvedené pouze u novějších záznamů tedy ve zhruba polovině případů.
- **Věk v letech, měsících a dnech** – jsou rozdělené do tří samostatných polí uvedené vždy v číselné podobě. Věk v letech je uveden u většiny osob starších jednoho roku. Věk v měsících a ve dnech se používá pro děti do jednoho roku, ale věk v měsících se může objevit i u starších dětí.

Matriky oddaných

Ideální záznam z matriky oddaných poskytuje tyto údaje:

- **Datum oddání** – je rozdělené do tří samostatných polí, které obsahují den, měsíc a rok uvedený vždy v číselné podobě.
- **Obec a číslo domu ženicha** – jsou rozdělené do dvou samostatných polí. Obec bývá uvedena zkratkou (celým jménem jsou napsány obce, které se vyskytují vzácně) a je u zhruba poloviny záznamů (není ovlivněno rokem pořízení záznamu) a číslo domu je v číselné podobě a uvedené také ve zhruba polovině případů.
- **Jméno, příjmení a povolání ženicha** – jsou rozdělené do tří samostatných polí a uvedené jsou vždy v řetězcové podobě. Povolání se vyskytuje jen u některých novějších záznamů.
- **Jméno, povolání a obec otce ženicha** – jsou rozdělené do tří samostatných polí a uvedené jsou vždy v řetězcové podobě. Jméno se nachází u většiny záznamů, povolání většinou u nových záznamů, obec pak jen v málo případech.
- **Jméno a rodné příjmení matky ženicha** – jsou rozdělené do dvou samostatných polí a uvedené jsou vždy v řetězcové podobě. Oba údaje se vyskytují pouze u nových záznamů od roku 1824.

- **Jméno, povolání a obec otce matky ženicha** – jsou rozdělené do tří samostatných polí a uvedené jsou vždy v řetězcové podobě. Tyto údaje jsou uvedeny pouze u záznamů, kde jsou údaje o matce.
- **Věk ženicha** – je číselný údaj v jednom samostatném poli. Tento údaj bývá pravidelně v záznamech od roku 1784.
- **Obec a číslo domu nevěsty** – jsou rozdělené do dvou samostatných polí. Údaje jsou uvedeny v podobném formátu a četnosti jako u ženicha.
- **Jméno, rodné příjmení nevěsty** – jsou rozdělené do dvou samostatných polí a uvedené jsou vždy v řetězcové podobě.
- **Jméno a povolání otce nevěsty** – jsou rozdělené do dvou samostatných polí a uvedené jsou vždy v řetězcové podobě. Jméno se nachází u většiny záznamů, povolání většinou u nových záznamů.
- **Jméno, příjmení a povolání předchozího manžela nevěsty** – jsou rozdělené do dvou samostatných polí a uvedené jsou vždy v řetězcové podobě. Tyto záznamy jsou uvedené, pokud je daná nevěsta vdovou. V tomto případě nejsou uvedeny předchozí údaje o otci a rodném příjmení nevěsty.
- **Věk nevěsty** – je číselný údaj v jednom samostatném poli. Tento údaj bývá pravidelně v záznamech od roku 1784 podobně jako pro ženicha.
- **Jméno a rodné příjmení matky nevěsty** – jsou rozdělené do dvou samostatných polí a uvedené jsou vždy v řetězcové podobě. Oba údaje se vyskytují pouze u nových záznamů od roku 1824 podobně jako pro ženicha.
- **Jméno, povolání a obec otce matky nevěsty** – jsou rozdělené do tří samostatných polí a uvedené jsou vždy v řetězcové podobě. Tyto údaje jsou uvedeny pouze u záznamů, kde jsou údaje o matce.