Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5212-92018

Matej Glemba

Knižnice a ich dáta pod lupou

Bakalárska práca

Vedúci práce: Ing. Nadežda Andrejčíková, PhD.

Máj 2021

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5212-92018

Matej Glemba

Knižnice a ich dáta pod lupou

Bakalárska práca

Študijný program: Informatika

Študijný odbor: Informatika

Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU, Bratislava

Vedúci práce: Ing. Nadežda Andrejčíková, PhD.

Máj 2021

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

ZADANIE BAKALÁRSKEHO PROJEKTU

Meno študenta:	**Glemba Matej**
Študijný odbor:	Informatika
Študijný program:	Informatika
Názov projektu:	**Knižnice a ich dáta pod lupou**

Zadanie:
Knižnice, podobne ako aj iné pamäťové a fondové inštitúcie, v svojich informačných systémoch uchovávajú množstvo rôznych typov dát. Jedná sa predovšetkým o dáta, ktoré vznikajú ako výsledok ich hlavnej pracovnej činnosti, teda dáta popisujúce dokumenty a ďalšie zdroje poznania, či transakcie vznikajúce pri poskytovaní služieb používateľom. Druhú skupinu dát tvoria dáta, ktoré sú z väčšej časti generované systémom a dokumentujúce jednotlivé procesy, či tvoriace základ pre štatistiky, ktoré knižnice využívajú pri rozhodovaní a riadení. Manažment knižníc však často potrebuje pre svoje rozhodovanie rôzne informácie a odpovede na viaceré otázky, ktoré sú v týchto dátach ukryté. Taktiež zmena prostredia má za následok zmeny v požiadavkách a potrebách používateľov a preto väčšina dnešných používateľov skôr ako číselné výstupy preferuje vizualizáciu týchto dát. Analyzujte preto dostupné štatistické metódy ako aj dáta v knižniciach, pričom sa zamerajte na metódy, ktoré umožňujú odhaľovať nové poznatky v dátach a existujúce možnosti využitia nástrojov pre tvorbu štatistík a výstupov súčasných informačných systémoch pre knižnice.
Na základe výsledkov analýzy navrhnite informačný systém, ktorý bude knižniciam poskytovať variabilitu pri tvorbe výstupných zostáv, štatistík a rôznych pohľadov na dáta z ich informačných systémov so zameraním na odhaľovanie nových poznatkov. Toto riešenie vhodnými technologickými prostriedkami implementujte a po otestovaní funkčnosti overte a vyhodnoťte správnosť tohto riešenia. Možnosti Vášho riešenia porovnajte s existujúcimi nástrojmi informačných systémov pre knižnice u nás a zahraničí.

Práca musí obsahovať:

Anotáciu v slovenskom a anglickom jazyku
Analýzu problému
Opis riešenia
Zhodnotenie
Technickú dokumentáciu
Zoznam použitej literatúry
Elektronické médium obsahujúce vytvorený produkt spolu s dokumentáciou

Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU, Bratislava; Cosmotron Slovakia, s r.o. Kopčany
Vedúci projektu: Ing. Nadežda Andrejčíková, PhD.

Termín odovzdania práce v letnom semestri : 17.5.2021

Bratislava 15.2.2021

doc. Ing. Valentino Vranić, PhD.
riaditeľ ÚISI

Čestné prehlásenie

Čestne vyhlasujem, že som túto prácu vypracoval samostatne, na základe konzultácií a s použitím uvedenej literatúry.

V Bratislave, Máj 2021 .....

. Matej Glemba

Poďakovanie

Poďakovanie pre Ing. Nadeždu Andrejčíkovú, PhD. za je trpezlivý prístup a veľkú ochotu pri konzultáciach.

Anotácia

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Študijný program: Informatika

Autor: Matej Glemba
Diplomová práca: Knižnice a ich dáta pod lupou
Vedúci projektu: Ing. Nadežda Andrejčíková, PhD.
Máj 2021

Cieľom tejto bakalárskej práce je navrhnúť knižničný informačný systém, ktorý bude slúžiť knižniciam na rôzny pohľad na knižničné dáta a ich využitie. V súčasnosti existuje niekoľko informačných systémov, ktoré okrem vyhľadávania titulov poskytujú aj rôzne štatistické výstupy z ich dát, ktoré im slúžia na rozhodovanie sa a riadenie knižnice. Práve dáta z knižníc sú hlavným problémom, nielen pre ich množstvo, ale aj diverzitu. Okrem knižných titulov si knižnica ukladá dáta generované systémom, transakcie pri poskytovaní služieb a všetky dáta vznikajúce ako dôsledok pracovnej činnosti. Pre prácu s knižničnými dátami sa používajú moderné prístupy dátovej analýzy ako princíp zhlukovania a sekvenovania spolu s algoritmami strojového učenia ako napríklad učenie s učiteľom. Použitím práve týchto aplikácií hĺbkovej analýzy na knižničných dátach chceme nájsť nové pohľady, hľadať nové vzťahy a súvislosti a tým vytvárať výstupy, ktoré vo výsledku vedia knižnici dať odpovede na otázky slabej predajnosti niektorých knižných titulov alebo naopak identifikovať frekventovane vypožičiavané tituly a tým zabezpečiť znásobenie kópií, prípadne poskytnúť podklady na odporúčanie pre špecifických zákazníkov na základe ich osobného profilu a obsahu výpožičiek.

Annotation

Slovak University of Technology Bratislava
Faculty of Informatics and Information technologies
Study program: Informatika

Name: Matej Glemba
Bachelor thesis: Knižnice a ich dáta pod lupou
Supervisor: Ing. Nadežda Andrejčíková, PhD.
Máj 2021

The aim of this bachelor thesis is to design a library information system that will serve libraries for different point of view of library data and their use. At present, there are several information systems that, in addition to searching for titles, also provide various statistical outputs from their data to help them decide and manage the library. Library data is the main problem, not only for its quantity but also for diversity. In addition to library titles, the library stores system-generated data, service delivery transactions and all data resulting from work activity. Modern data analysis approaches are used to work with library data as a clustering principle or the use of sequence patterns together with machine learning algorithms such as supervised learning. Using these data mining applications on library data, we want to find new insights, find new relationships and contexts and thereby produce outputs that ultimately lead the library to answer questions about poor marketability of some books or identify frequent borrowed titles to ensure multiple copies or provide referrals to specific customers based on their personal profile and rental content.

Obsah

Slovník pojmov

ISBD Medzinárodný štandardný bibliografický popis

Z3950 Protokol pre vyhľadávanie v textových databázach

DDC Deweyova desatinná klasifikácia

LLC Knižnica kongresovej klasifikácie

OPAC Online katalóg

IPAC Online katalóg pre ARL

IS Informačný systém

KIS Knižnično-informačný systém

REST Representational state transfer

SOAP Simple Object Access Protocol

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting

MARC Machine readable cataloging

DBMS Databázový systém

MDT Medzinárodné desatinné triedenie

MARCXML Syntaxová XML reprezentácia MARC 21

MODS Metadata object description schema

MADS Metadata authority description schema

Zoznam obrázkov

Kapitola 1

ÚVOD

Knižnično-informačné systémy pracujú s nespočetným množstvom dát, ktoré musia spracovávať, evidovať, efektívne s nimi manipulovať a tým zabezpečiť ich najlepšie využitie pre potreby knižnice. Práve manipulácia s dátami a ich efektívne použitie pre prosperitu knižnice je problematika, ktorou sa zaoberám v mojej bakalárskej práci.

V nasledujúcich kapitolách zanalyzujem aktuálny stav knižničných systémov, zadefinujem s akými typmi dát prichádzajú do kontaktu, akým spôsob dané dáta identifikujem, zatriedim podľa zámeru ich využitia a následne na ne aplikujem konkrétne štatistické metódy použitím dostupných štatistických nástrojov na vytvorenie štatistických výstupov, s ktorými bude môcť zamestnanec knižnice pracovať a odhaľovať tak nové poznatky a vzťahy. Hlavným cieľom bakalárskej práce je navrhnúť knižnično-informačný systém, ktorý bude knižniciam poskytovať variabilitu pri tvorbe výstupných zostáv, štatistík a rôznych pohľadov na dáta. V tejto práci sa zameriam najmä na dáta knižníc a spôsoby ich získavania, tak aby sme dané dáta mohli analyzovať a poskytovať používateľom pohľad na tieto ich dáta z iného uhla pohľadu.

Kapitola 2

Kultúrne dedičstvo a pamäťové inštitúcie

Kultúrne dedičstvo svojou kvantitou a diverzitou je veľmi obsiahle, kvôli čomu sa kategorizuje na dve podčasti, hmotné a nehmotné [31].

2.1 Nehmotné dedičstvo

Jeho hlavnou charakteristikou je nehmotná podstata. Každá krajina disponuje reprezentatívnym zoznamom nehmotného kultúrneho dedičstva. Napríklad v rámci Slovenskej republiky, bolo najnovším prvkom nehmotného dedičstva v roku 2019 pridané drotárstvo [1]. Prvky, ktoré sú zapísané v nehmotnom dedičstve krajiny následne môžu byť po schválení zapísané vo svetovom nehmotnom kultúrnom dedičstve UNESCO.

2.2 Hmotné dedičstvo

Druhou podčasťou je hmotné kultúrne dedičstvo. Analogicky je charakterizované bázou, ktorá je hmotná, z toho vyplýva, že prvok hmotného dedičstva je výsledkom ľudskej činnosti. Hmotné dedičstvo je všeobecne rozdelené do štyroch fondov: pamiatkového, zbierkového, archívneho a knižného [43].

2.2.1 Fondy hmotného dedičstva

Pamiatkový fond na Slovensku pozostáva z národných kultúrnych pamiatok, pamiatkových rezervácií, pamiatkových zón a lokalít, ktoré sa nachádzajú v zozname svetového dedičstva UNESCO. Archívny fond je štruktúra zbierok, archívnych dokumentov ako aj audiovizuálnych dokumentov, ktoré sú umiestnené v sieti archívov. Za archívný dokument v rámci Slovenskej republiky môžme všeobecne považovať záznam, ktorý má trvalú dokázateľnú a historickú hodnotu pre poznanie slovenských dejín *(tzv. archívne minimum)*. Príklady archívnych dokumentov nejakej organizácie sú napríklad správy a analýzy činnosti organizácie, správy z vedeckovýskumných úloh, účtovné doklady, štatistické výkazy, ale aj rôzne publikácie ako aj audiovizuálne materíaly. Druhým fondom nehmotného kultúrneho dedičstva je zbierkovy fond, ktorý zahŕňa obsah múzeí a galérií. Z hľadiska hierarchie štruktúr je zbierkový fond považovaný za najväčšiu štruktúru zbierkového systému. Najväčším zbierkovým fondom na území Slovenskej republiky je Zbierkový fond Slovenského národného múzea, kde je aktuálne evidovaných celkovo 4 milióny zbierkových predmetov [43, 34, 44].

2.2.1.1 Knižničný fond

Posledným z fondov nehmotného kultúrneho dedičstva je knižničný fond. Pamäťovou inštitúciou, ktorej sa knižničný fond týka je knižnica a všetky jej doku-

menty, ktoré eviduje, uskladňuje, doplňuje a sprístupňuje medzi čitateľov. Definíciou knižničného fondu podľa Zákona č. 126/2015 Z. z. článku 3 je „súbor všetkých knižničných dokumentov, účelovo vybraných, sústavne doplňovaných, uchovávaných, ochraňovaných a sprístupňovaných používateľovi [2].“ Ako všetky predošlé nehmotné fondy kultúrneho dedičstva aj knižničný fond vieme charakterizovať určitými vlastnosťami. Okrem toho, že je funkčný, to znamená, že jeho reprezentácia, čiže knižnica spĺňa zákonom ustanovené náležitosti funkcie knižnice, je aj organizovaný a výberový, čo naznačuje, že čitateľ je schopný vyhľadať každý dokument uchovaný v knižnici, ktorý sa dostal do knižničného fondu podľa určitých ustanovených a preddefinovaných kritérií. Poslednou charakteristikou je fyzická dostupnosť knižničného fondu. Táto vlastnosť hovorí o fyzickom sprístupnení knižničných dokumentov, či už textových, zvukových alebo obrazových. Konkrétne sa jedná napríklad o knihy, rukopisy, grafiky, mapy, kartografie, patentované dokumenty, technické normy, obrazy, kresby, ale aj audiovizuálne dokumenty a mnohé iné. Okrem toho sú súčasťou fondu aj prezenčné a absenčné výpožičky. V rámci Slovenskej republiky medzi univerzitnými knižnicami, má napríklad knižničný fond Slovenskej ekonomickej univerzity viac ako 360 000 publikácii. Univerzitná knižnica Konštantína filozofa v Nitre disponuje viac ako 300 000 knižničnými jednotkami [44].

Kapitola 3

Knižnice a knižničné systémy

3.1 Knižnica

Knižničný fond je teda reprezentovaný pamäťovými inštitúciami, knižnicami, ktoré vytvárajú dokopy vyššiu štruktúru a to knižničný systén. U nás sa nazýva Knižničný systém Slovenskej republiky, ktorý definuje zákon číslo 126/2015 Z. z. a právne spadá pod Ministerstvo kultúry Slovenskej republiky. Do tohto knižničného systému okrem Slovenskej národnej knižnice, ktorá koordinuje celkový rozvoj knižničného systému, spadajú vedecké, akademické, verejné, školské alebo aj špeciálne knižnice. Sumárne disponuje viac ako 6,5 miliónami knižničných a špeciálnych dokumentov [39]. Knižnica je základná jednotka knižničného systému, ktorej primárnou úlohou je evidencia dokumentov. Podľa toho aké ma daná knižnica zameranie, či pôsobnosť ju vieme charakterizovať ako informačnú, vzdelávaciu alebo kultúrnu inštitúciu. Jej hlavnou úlohou je knižničné dáta vyhľadávať, získavať, spracovávať, uchovávať a sprístupňovať [**wikipedia_2019lib**]. V rámci Slovenska podobne ako aj v iných krajinách sveta sú knižnice organizované a riadené systémom knižníc na rôznych úrovniach: inštitucionálnej, oborovej, regionálnej, národnej a medzi-

národnej. Keďže rôzne typy knižníc častokrát spracovávajú rovnaké dokumenty, snažia sa o efektívny spôsob zdielania údajov. Ešte pred automatizáciou knižníc, si samotné knižnice zadefinovali štandardy spracovávania údajov o dokumentoch s cieľom spracovať obsah jednotlivých dokumentov vrátane základných opisných metadát. Jednou z prvých noriem bola norma ISBD *(z angl. International Standard Bibliographic Description)*. Dôležitým prelomom bol protokol Z3950, ktorý neskôr vyšiel ako ISO23950 norma a ANSI norma. Výhodou tohto protokolu bol fakt, že knižničné systémy v rámci svojej komunikácie nepotrebovali vedieť syntax alebo sémantiku spracovávania údajov ostatných systémov. Cieľom protokolu bolo vyhľadávanie, prehľadávanie, triedenie dát ako aj komunikácia s databázou [3].

3.1.1 Klasifikačný systém v knižniciach

Pri klasifikáci, usporiadávaní a uskladňovaní všetkých knižničných dokumentov v rámci knižnice je potrebné dodržiavať klasifikačný systém, aby bolo možné jednotlivé knižničné dokumenty rýchlo a efektívne nájsť. Klasifikačný systém usporadúva dokumenty vo viacerých vrstvách. Keďže existuje množstvo klasifikačných systémov a ďalšie sa stále vyvýjajú, je potrebné ich nejakým spôsobom rozdeliť, napríklad podľa spôsobu používania. Univerzálne schémy, kde sú začlenené napríklad Deweyova desatinna klasifikácia a univerzálna desatinná klasifikácia. Špecifické klasifikačné schémy zahŕňajú jednotlivé typy subjektov, ako NLM klasifikácia alebo Dickinson klasifikácia. Posledným typom sú národné schémy, ktoré už z názvu hovoria o špecifickéj orientácií na krajiny.

Iným kritériom delenia je funkcionalita, kde vieme zatriediť klasifikačné systémy ako enumeratívne, hierarchické alebo analyticko-syntetické. Enumeratívny spôsob hovorí o abecednom zoradení nadpisov subjektov, ktorým je priradené číslo. Hierar-

chický spôsob rozdeľuje subjekty od najšpecifickejších po najvšeobecnejšie. Veľmi zriedkavo sa však stane, že sa využíva iba jeden spôsob, čiže kvôli väčšej efektivite je lepšou možnosťou zmes viacerých spôsobov s prízuvkom na jeden konkrétny. V súčastnosti sú najbežnejšie klasifikačné systémy DDC (*Deweyova desatinna klasifikacia*) a LLC (*Knižnica kongresovej klasifikácie*), ktoré sú primárne enumeratívne. Jednotlivé klasifikačné systémy, slúžia teda knihovníkom na vytváranie a používanie konštrukcií, ktoré slúžia ako nástroje pre knižnice. Obsahuju hlavné katalógy, doménové katalógy, indexy, jedinečné identifikátory a jedinečné identifikačné tokeny ako aj rôzne artefakty. Hlavný katalóg sa väčšinou nachádza v blízkosti vstupu do knižnice. Je to hlavné miesto odkial sa čitateľ vie nasmerovať na konkrétnu časť knižnice, ktorá je určená nejakým konkrétnym predmetom ako história, vzdelanie, beletria, kde by mal nájsť už konkrétny tematický katalóg. Hlavný katalóg teda pozostáva zo všetkých tematických katalógov. Tematické katalógy analogicky obsahujú už informácie a odkazy na konkrétne artefakty, ktoré spadajú pod danú tému. Z praktického hľadiska, len knižnice väčších rozmerov a väčšieho množstva knižničných dokumentov obsahujú doménové katalógy, lebo v prípade velkosťou menších knižníc dokáže hlavný katalóg obsiahnuť všetky potrebné informácie a odkazy, bez toho aby to už nebolo náročné pre čitateľa sa v hlavnom katalógu vyznať. Hierarchicky ďalším nástrojom sú indexy. Indexy zoskupujú jednotlivé dokumenty podľa podstatných faktorov zužovania obsahu. Príkladom môžu byť indexy podľa autora, podľa vydavateľstva, podľa roku vydania, podľa témy a tak ďalej. Posledným nástrojom, ktorý sa viaže už na konkrétne dokumenty sú jedinečné identifikátory, ktoré priradzujú artefaktom jedinečné číselné reťazce [38]. Dnes sú tieto kartotéky v kamenných obchodoch nahradené tzv. OPAC, alebo v ARL používaným IPAC, ktorý ponúka vyhľadávanie podľa viacerých selekčných kritérií, či kombináciou viacerých termínov. Takýto spôsob vyhľadávania je dnes už priamo integrovaný vo webových stránkach knižníc, alebo aj v iných IS, ktoré

k tomu môžu zväčša využitť API daného systému v podobe REST alebo SOAP služieb, ktoré sa stali bežnou súčasťou KIS. Špecifickú skupinu tvoria tezaury, slovníky, taxonómie a iné klasifikačné systémy, ktoré sa používajú pre vyjadrenie obsahu a správne zaindexovanie dokumentov v knižniciach. Z tezaurov sa u nás využíva predovšetkým MESH. Spracovávaný a prevádzkovaný na národnej úrovni v systéme ARL , ktorý používa Slovenská lekárska knižnica, ktorá je zodpovedná za slovenskú verziu tohoto medzinárodného tezauru. Ďalej je to čiastočne Agrovoc, Eurovoc, AAT a iné. Z uvedeného je vidno, že ide o špecializované oborové tezaury. Preto sú zaujímavejšie pre nás indexačné systémy, ktoré majú svoju vlastnú hierarchiu a dnes sú aj vzájomne kompatibilné.

3.2 Knižničný informačný systém

Všetdy organizačné procesy knižnice ako akvizícia, cirkulácia, klasifikácia, katalogizácia, technická podpora tvoria všeobecne vyšší abstrakčný celok nazývaný systém správy knižníc. V prípade, že inštitúcia nedisponuje informačným systémom, systém správy knižníc je chápaný ako jednotlivé pracoviská alebo oddelenia knižnice. V poslednej dobe sa rozširuje pojem Digitálna knižnica. Digitalizácia je proces, kedy sa fyzické dokumenty transforumujú to digitálnej formy. Výhodou je, že jeden exemplár v digitálnej forme môže zdielať viacero používateľov. Druhou výhodou je zachránenie fyzických dokumentov, ktoré napríklad môžu mať už vyšší vek a ich častým používaním by viac strácali na kvalite. Digitalizácia je teda separátny proces, pri ktorom nemusí fyzická verzia dokumentu zaniknúť. Eliminuje hlavný problém tradičných knižníc, čo je fyzická kapacita respektíve úložný priestor, rovnako aj zmenšuje náklady na údržbu. Digitálna forma knižnice poskytuje služby nad digitálnym fondom. Systém zabezpečuje aby jednotlivé dokumenty boli uložené v súlade s medzinárodnými štandardmi, čo podporuje vyššiu mieru inte-

roperability a tiež nezávislosť od IS [28, 4, 29]. KIS rovanko, ako aj iné systémy pozostáva z viacerých celkov, ktoré postupne automatizujú procesy knižnice od objavovania, získavania, cez spracovanie , archivovanie, sprístupňovanie, zdieľanie, ale aj digitalizáciu dokumentov, ich poskytovanie, ako aj ošetrovanie a uchovávanie a to všetko s ohľadom na kontrolu a riadenie práv prístupu k chráneným, či licencovaným zdrojom [28, 4, 5]. Ako bolo už spomenuté v predošlej podkapitole, prístup k dátam zabezpečuje protokol z39.50. Avšak v súčastnosti je to práve OAI-PMH štandard, ktorý sa používa v rámci OPAC modulu knižnice a ktorý definuje spôsob ako poskytovateľ dát poskytuje svoje dáta pre iné systémy za účelom zdielania dát alebo za účelom vzájomnej kooperacie a poskytovaní rozšírených služieb. Jednotlivé IS môžu používať vlastný BATH profil, ktorý definuje syntax a sémantiku jazyka komunikácie medzi IS.

3.2.1 Typy knižničných systémov na Slovensku

Jednými z najpoužívanejších knižničných informačných systémov v rámci akademických ale aj verejných knižníc na území Slovenskej republiky sú systémy DAWINCI, Clavius, ARL a KIS MaSK. Systém DAWINCI vyniká hlavne integrovaním všetkých najnovších prvkov z oblasti manažmentu, server-cliet aplikácií a relačných databázových systémov do jedného komplexného knižničného informačného systému. Zo všeobecne známych výmenných formátov podporuje formát MARC, normu ISO 2709 a protokoly Z39.50 a OAI-PMH. Okrem toho je definovaných viacero štandardov pre zápis syntaxe týchto dát, ale tomu musia predchádzať pravidlá, ako tieto dáta zapisovať do týchto schém. Najskôr s nástupom UNIMARCu sa u nás aplikovali Angloamerické pravidlá katagolizácie (AACR, AACR2, AACR2R), tieto však na súčastnú dobu digitalizácie sú už zastaralé a tak ako nástupca AACR2 prišiel RDA (Resource Description and Access). Podstatou RDA je umožnenie prístupu k čo najviac dátam pre koncového používateľa,

ktorý si vyhľadáva jednotlivé tituly. Oproti AACR2 má RDA niekoľko vylepšení ako napríklad neskracovanie bibliografických údajov, ale ukladanie v celej forme. RDA je kompatibilné so všetkými medzinárodnými modelmi a normami. RDA ako webový toolkit je založený na koncepčných modeloch FRBR (požiadavky na bibliografické záznamy), FRAD (požiadavky na údaje o autorite) a FRSAD (požiadavky na údaje o autorite subjektu) [6, 7]. Druhým vhodnýnm príkladom, môže byť KIS Advanced Rapid Library [28, 4]. Systém je vhodný pre všetky typy knižníc, bez rozdielu veľkosti fondu, zamerania knižnice, či počtu používateľov , množstvu transakcii a podobne. Systém plne podporuje prácu s full-textovými, zvukovými, aj audiovizuálnymi a multimediálnymi dokumentami. Systém je implementovaný na rovnomennej platforme , ktorá umožňuje nakonfigurovať a poskladať si systém presne na mieru podľa požiadaviek danej inštitúcie. Na rozdieľ od všetkých ostatných u nás využívaných systémov v knižniciach, nepracuje priamo v relačnom databázovom systéme, ale v systéme je priamo zapuzdrená viacrozmerná databáza Cache s možnosťou relačného prístupu k dátam v nej uložených. To má výhodu v tom, že pre prístup k dátam a prácu s nimi nie je vyžadovaná znalosť DBMS, čo umožnuje, že knihovníci sú priamo administratormi databáz a môžu si robiť ľubovolné výstupy, pohľady na dáta, či modifikovať a rozširovať datový model.

Kapitola 4

Typy knižničných dát a štandardov

Okrem samotných knižničných dokumentov knižnica disponuje aj inými dátami, ktoré sú kľúčové pre KIS. Prvým typom sú dáta vložené používateľom. Tieto dáta už zo samotnej povahy práce knižníc môžeme rozdeliť na 4 hlavné skupiny. Dáta opisujúce samotné tituly, dáta opisujúce exempláre dokumentov, údaje o používateľoch, dáta o transakciách. Medzi transakcie patria výpožičky, návraty, upomienky, žiadanky, rezervácie, platby. Druhým typom dát sú dáta generované samotnými úkonmi IS. Jedná sa o dáta, ktoré hovoria o tom, čo robí samotný softvér KIS, čo robí používateľ, napríklad aké sú jeho žiadosti a výstupy, čo a kedy vyhľadáva, čo našiel a čo nenašiel, čo si stiahol a podobne. Ak hovoríme o dátach v knižniciach, nemôžeme zabúdať ani na tie, ktoré nám umožňujú vyhľadávať a zdieľať dáta naprieč IS. Tu hovoríme o dátach zaznamenávaných v súboroch autorít. Podľa Vodičkovej autority predstavujú unifikované selekčné prvky s nevyhnutným odkazovým a poznámkovým aparátom [8]. Z hľadiska informatiky a práce s jazykom sa to môže chápať ako termíny, ktoré môžu byť aj viacslovné, ale nesmú to byť

homonymá, spolu s prepojeným na všetky súvisiace termíny (nadradené, podradené, asociatívne) ako aj všetky synonymá, akronymi a iné altrernatívne spôsoby vyjadrenia toho istého významu ako má selekčný termín.

4.1 Zatriedenie dokumentov do oborov

Knižničné dokumenty sa z hľadiska obsahu dokumentu triedia do 26 oborov pomocou metódy konspektu. Tie sú potom ďalej členené podľa potrieb v rámci daného odboru na konkrétnejšie skupiny konspektu, pričom kategórie aj skupiny konspektu vieme vyjadriť kombináciou a rozpatím MDT. Napríklad História má cez 80 konkrétnejších skupín konspektu (Svetové dejiny, Európske dejiny, Pravek, ...). Poslednou jednotkou v systéme, ktorú definuje priamo MDT je konkrétny titul, o ktorom vieme jeho názov, autora, spôsob nadobudnutia (kúpa, dar, výmena), lokáciu a dislokáciu, signatúru alebo čiarový kód a ďalšie údaje, ktoré sú popísané v Marc 21 schémach. Môže to byť však knižná edícia, prípadne viaczväzkový dokument. Samotný čitateľ sa ale dostáva do kontaktu s konkrétnym tlačeným exemplárom alebo knižnou jednotkou.

4.2 Dáta o užívateľoch

Knižnice okrem dokumentov evidujú aj údaje o svojich čitateľoch, ktoré môže knižnica spracovávať vo svojich IS. Pokiaľ sa čitateľ aktívne zapojí do výpožičkového procesu, prípadne sa zaregistruje ako člen knižnice, knižnica si ukladá jeho osobné údaje pre svoje vlastné potreby. Údaje o čitateľoch knižnice definuje zákon o ochrane osobných údajov [9]. Knižnica nemá vekové obmedzenie pre svojich čitateľov, avšak ak užívateľ nemá viac ako 15 rokov, zastupuje ho jeho zákonný zástupca. Údaje o čitateľovi sú nasledovné:

- Meno a priezvisko
- Dátum narodenia
- Adresa bydliska - trvalé alebo prechodné
- Tel. číslo, e-mail
- špecifické údaje ako (ISIC alebo dosiahnuté vzdelanie)

4.3 Dáta o exemplároch

Podstatným typom sú dáta všetkých knižničných jednotiek dokumentov knižničného fondu, ktorým predchádzajú informácie o ich tituloch. Presnú štruktúru dát o knižničných jednotkách opisuje vyhláška Ministerstva kultúry Slovenskej republiky č. 201/2016 Z. z. [10]. Knižnica eviduje dokumenty v prírastkovom a úbytkovom zozname. Každý prírastkový záznam o knižničnom dokumente obsahuje:

- prírastkové číslo - je jedinečné
- signatúru, ak sa nezhoduje s prírastkovým číslom
- meno a priezvisko autora a názov
- vydavateľské údaje
  - miesto vydania
  - označenie vydavateľa
  - rok vydania
- cenu
- spôsob nadobudnutia
- dátum zápisu

- medzinárodné štandardné číslo knihy (ISBN)
- jazyk knižničného dokumentu
- údaje o prílohách
- poznámku o vyradení

Úbytkový záznam obsahuje :

- poradové číslo
- prírastkové čiślo
- signatúru, ak sa nezhoduje s prírastkovým číslom
- meno a priezvisko autora a názov
- dátum zápisu do zoznamu úbytkov
- dôvod vyradenia s odkazom na doklad o vyradení
- cenu [10].

4.4 Dáta z transakcií

Dáta z transakcií pozostávajú z platieb, návratov, upomienok, rezervácií, žiadaniek a výpožičiek. Jednou z hlavných služieb knižnice je výpožičková služba, ktorá okrem evidenčných údajov o knižničnom dokumente obsahuje aj údaje o používateľovi knižničného systému. Ak je používateľ mladší, knižničný systém eviduje aj osobné údaje jeho zákonného zástupcu. Po registrácií užívateľom v systéme mu je pridelené jedinečné identifikačné čislo. Všetky výpožičky si systém eviduje vo vlastnom zozname výpožičiek. Záznam z evidencie výpožičiek podla výpožičkového poriadku z knižnice v Topolčanoch obsahuje:

- meno a priezvisko používateľa
- číslo čitateľského preukazu alebo identifikačnej karty používateľa
- prírastkové číslo alebo signatúru, ak sa nezhoduje s prírastkovým číslom alebo identifikačný kód knižničného dokumentu
- dátum výpožičky a dátum vrátenia vypožičaného knižničného dokumentu [11].

Výpožičky môžu byť absenčné alebo prezenčné. Konkrétny čitateľ si väčšinou vie z konkrétneho titulu vypožičať iba jeden exemplár na dobu, ktorá sa môže v jednotlivých knižniciach líšiť. Takisto sa vie líšiť aj maximálne množstvo vypožičaných dokumentov, zvyčajne však 10. Ak si čitateľ rezervuje knižný dokument, zvyčajne má určitú lehotu na jeho prevzatie. Počas tejto doby je daný exemplár v stave držanej rezervácie. V prípade, že si čitateľ neprevezme svoj rezervovaný dokument a je o ten daný titul záujem, čiže systém eviduje žiadosti o rezervácie, tak v zozname čakateľov sa podľa dátumu vzniku podania žiadosti o rezerváciu, pridelí ďaľšiemu čitateľovi, ktorému následne rovnako plynie lehota na prevzatie. Ak o daný titul nie je záujem, tak sa v systéme uvoľní pre následne vypožičanie. Výpožičky môžu byť medziknižničné, ale aj medzinárodné medziknižničné výpožičky. Pri takýchto výpožičkách sa využívajú žiadanky. Žiadanka na konktrétny titul je alternatívou k objednaniu titulu prostredníctvom online modulu OPAC-u. Pri predĺžení výpožičnej lehoty dokumentu je zaslaná čitateľovi upomienka, ktorá má viacero stupňov [12].

Podľa pripraveného datasetu [13] si vieme rozobrať konkrétne dáta, ktoré si knižnica (v Českej republike) eviduje o svojih transakciách:

- Identifikátor transakčného dokladu
- Dátum vytvorenia dokladu

- Poradie operacií v rámci dokladu
- Typ transakcie (Výpožička, Návrat, Prolongácia, Upomienka, Žiadanka, Nevyžiadaná žiadanka, Rezervácia, Platba, Presun, Štatistika, Voľný výber, ...)
- Dátum transakcie
- Pobočka
- Identifikácia holdingu
- Identifikácia katalógového záznamu
- ISBN
- Knižná väzba
- Cena
- Identifikátor českej národnej knižnice
- Externý identifikátor
- MDT
- Anonymizovanie dokladu
- Pohlavie
- Hash používateľa
- Typ používateľa
- PSČ
- Vek
- Dĺžka vypožičania (od vypožičania po návrat)
- Dĺžka transakcie (od vytvorenia transakcie po poslednú transakciu)

- Počet prolongácií za danú transakciu
- Počet upomienok za danú transakciu

Na tomto datasete vidíme, že v prvej časti sú tam informácie o danej transakcií a operáciach, nasledujú informácie o knižných dokumentoch ako ISBN, Cena, MDT, Identifikátory až nakoniec údaje o používateľovi ale to iba v prípade, že daný transakčný doklad nie je anonymizovaný.

4.5 Štatistické dáta

Knižnica na konci roka vydáva správu o činnosti knižnice za daný kalendárny rok podľa predlohy z Ministerstva Kultúry SK. Ročný výkaz sa skladá z nasledovných modulov:

- Modul knižničného fondu
- Modul výpožičky a služby
- Používatelia služieb knižnice, návštevníci, podujatia a edičná činnosť knižnice
- Ekonomické ukazovatele – finančné zabezpečenie činnosti knižnice
- Osoby zabezpečujúce činnosť knižnice

V tejto správe je štandardne zahrnutý prehľad knižničného fondu, to znamená aký je počet knižných jednotiek pre druh knižničného dokumentu, koľko nových titulov bolo pridaných, koľko vyradených a s tým spojená aj finančná stránka veci. Druhým typom dát v ročnej správe je počet registrovaných používateľov rozdelených podľa vekových kategórií a porovnanie prírastku a úbytku s predošlymi rokmi [14, 15]. Čo sa týka výpožičkových dát, správa zahŕňa koľko krát bola využíta výpožičková služba a koľko dokumentov bolo vypožičaných. Tieto dokumenty sa dajú rovnako rozdeliť napríklad podľa ich typu. Rovnako sa eviduje pokles prípadne nárast

oproti minulým rokom. Ako príklad by som použil štúdiu vykonanú na *University of Colorado*, kde sa robila štatistická analýza knižničných dát [35].

Rank	Conspectus Subject Category	Circulation Transactions	Circulating Items	Transactions per Item
1	Music	53,855	7,230	7.4
2	Computer Science	35,378	5,202	6.8
3	Sociology	106,724	17,809	6.0
4	Physical Education and Recreation	14,432	2,409	6.0
5	Art and Architecture	106,186	17,962	5.9
6	Anthropology	19,424	3,331	5.8
7	Psychology	31,281	5,376	5.8
8	Geography and Earth Sciences	31,473	5,552	5.7
9	Engineering and Technology	108,834	19,712	5.5
10	Mathematics	45,037	8,247	5.5
11	Performing Arts	25,106	4,661	5.4
12	Physical Sciences	52,423	9,786	5.4
13	Medicine	73,555	13,846	5.3
14	Agriculture	11,109	2,145	5.2
15	Biological Sciences	40,308	7,961	5.1
16	Language, Linguistics, and Literature	280,667	56,631	5.0
17	History and Auxiliary Sciences	230,262	46,515	5.0
18	Philosophy and Religion	91,324	18,696	4.9
19	Chemistry	13,011	2,775	4.7
20	Business and Economics	102,587	23,027	4.5
21	Political Science	52,108	11,745	4.4
22	Law	15,929	3,638	4.4
23	Invalid or unknown	44,707	10,419	4.3
24	Education	37,425	8,870	4.2
25	Library Science, Generalities, and Reference	15,595	4,972	3.1

Obr. 4.1: Priemerný počet výpožičiek podľa typu dokumentu [35]

Takými viac systémovými dátami môžu byť dáta priamo spojené s prevádzkou systému. Sú to dáta typu počet návštev systému, počet opakovaných návštev z rovnakej IP adresy, prípadne ak sú poskytnuté lokalizačné údaje tak z rovnakého miesta. Tieto dáta vedia napomôcť systému viac kategorizovať používateľa na aktívneho respektíve pravidelne sa zapájajúceho alebo na občasného respektíve náhodného. Pokiaľ by sme chceli vedieť počet návštev kamennej knižnice vedia nám napríklad napomôcť dáta z tretích strán ako Google, ktorý z lokalizačných

údajov používateľov vie určiť, kedy zažíva knižnica najväčší nápor. To môže dopomôcť napríklad k personálnemu posilneniu alebo v rámci systému prispôsobiť a zefektívniť časovú odozvu systému a predísť jeho padnutiu [16].

4.6 Štandard Marc

Reprezentácia dát knižničného fondu v digitalizovanej forme má svoje štandardy respektíve normy. Najviac používaným je štandard MARC. MARC určuje ako maju byť dáta reprezentované pre strojové čítanie a spracovanie. Keďže si každá krajina vytvárala vlastnú verziu MARC štandardu ako napríklad US MARC, CAN MARC, bola potrebná unifikácia týchto štandardov jednotlivých štátov aby bola zabezpečená bezproblémová komunikácia medzi knižnicami bez ohľadu na to, v ktorej krajine sa daný záznam vytvoril. Preto vznikol UNIMARC. Od roku 1993 má aj Slovenská republika slovenský preklad UNIMARC-u. Okrem UNIMARC štandardu sa v súčastnosti najmä používa verzia MARC 21, ktorá je spojením viacerých formátov a ktorá je prispôsobená 21. storočiu. Samotný MARC 21 štandard pokrýva 5 typov metadátových schém:

- 1. Schéma pre bibliografické údaje
- 2. Schéma pre autoritatívne údaje
- 3. Schéma pre holdingy
- 4. Schéma pre klasifikačné údaje
- 5. Schéma pre komunitné údaje [7, 17].

Formát MARC 21 pre bibliografické údaje s pravidlami označenia obsahu definuje kódy a konvencie (tagy, indikátory, kódy podpolí a kódované hodnoty), ktoré identifikujú údajové prvky v bibliografických záznamoch MARC. Bibliografický

záznam MARC obsahuje tieto komponenty:

- 1. Hlavička (návestie) - má pevnú dĺžku 24 znakov
- 2. Adresár - rad zápisov, každý má 12 znakov, obsahuje tag a dĺžku poľa so štartovacou pozíciou variabilného poľa v zázname. Duplicitné tagy sa odlišujú pozíciou polí v zázname.
- 3. Variabilné polia - každé je označené 3-znakovým kódom a každé ma kód konca poľa. Na konci je aj kód konca záznamu.
  - Variabilné riadiace polia - neobsahujú indikátory ani kódy podpolí, sú to polia s označením 00X.
  - Variabilné polia údajov - všetky ostatné variabilné polia. Obsahujú tagy uvedené v adresári, následne 2 znaky indikátora a 2 znaky kódu podpoľa. Variabilné polia údajov sú zoskupené v blokoch, pričom prvý znak 3-miestneho tagu hovorí o funkcionalite bloku.
    - 0xx - Riadiace údaje, kódované údaje, identifikačné čisla
    - 1xx - Hlavné záhlavie (okrem tohto bloku, aj bloky 4XX, 6XX, 7XX a 8XX maju podobnú štruktúru)
      - X00 - osobné mená
      - X10 - korporatívne mená
      - X11 - názvy zhromaždení
      - X30 - názové záhlavia
      - X40 - bibliografické názvy
      - X50 - tematické termíny

    - X51 - geografické názvy

* 2xx - Popisné údaje (názov, vydanie, poradie vydania, vydavateľstvo, ...)
* 3xx - Popis
* 4xx - Edícia
* 5xx - Poznámky
* 6xx - Polia prístupu podľa predmetu
* 7xx - Vedľajšie záhlavie a prepojovacie polia (iné vstupy ako podľa predmetu alebo edície)
* 8xx - Vedľajšie záhlavie pre edície alebo holdingy
* 9xx - Priestor pre lokálnu implementáciu

Variabilné polia obsahujú ako je vyššie spomenuté indikátory a kódy podpolí. Indikátory sú teda prvé dva znaky variabilného poľa. Sú interpretované nezávisle od ostatných indikátorov. Kódy podpolí sú rovnako dva znaky, ktoré oddelujú oddelovačmi (ASCII hodnota oddelovača je IF hex, respektíve znak doláru, za ktorým ide identifikačný znak) v poli tie údaje, s ktorými sa zrejme bude manipulovať. Doležitou vlastnosťou bibliografických záznamov je opakovanie polí a podpolí. Niektoré polia sa spravidla môžu opakovať avšak napríklad pole hlavného záhlavia je neopakovateľné a môže obsahovať iba jedno podpole "$a", ktoré hovorí o osobnom mene, ale môže obsahovať viac podpolí "$c", ktoré hovoria o titule. V prípade, že sa nejaké hodnoty kódov v zázname nepoužívajú rieši sa to výplňovým znakom, ktorý zvykne byť "[7].

4.6.1 Syntaxové reprezentácie

MARC 21 má viacero syntaxových reprezentácií ako MARCXML, MODS, MADS, ISO 2709. Metadata object description schema je XML schéma, ktorá podporuje správu a prístup k zdrojom a výmenu ich zakódovaných popisov, čo vie riešiť MADS schéma. MODS a MADS spolu úzko súvisia a je možné sa odkazovať z jednej schémy na druhú [18]. Okrem hlavných root elementov, ktoré zastrešujú jednotlivé záznamy *(element mods)* a list záznamov *(element modsCollection)*, tvoria podstatu MODS schémy nasledovné top elementy v rámci root elementu mods. Každý top element obsahuje svoje atribúty a sub-elementy. Všetky top elementy sú optional, avšak vždy tam musí byť aspoň jeden. Bližšie informácie k najaktuálnejšej verzii MODS 3.7 sú na oficiálnej stránke [19]. *Metadata authority description* schema je XML schéma, ktorá podporuje správu k dátam v popise zdroja. MADS schéma popisuje jednotlivé elementy v MODS schéme, to znamená, že napríklad top element titleInfo vieme popísať pomocou vlastnej MADS schémy a preto je možné sa vo všeobecnosti z MODS schémy odkazovať na MADS schému.

4.6.1.1 MARCXML

MARCXML je XML schéma MARC štandardu za ktorej vznikom môžeme vidieť zjednodušenie formátu MARC 21 pre lepšiu agregáciu, čitateľnejšiu reprezentáciu alebo aj kontrolu správnosti na základe XML overovacích nástrojov. Na rozdieľ od predošlých schém, MARCXML je všeobecnejšia verzia MODS a MADS, preto je veľmi časté, že sa konvertujú dáta medzi týmito schémami. Základnou štruktúrov formátu MARCXML je XSD schéma, ktorá obsahuje 2 hlavné elementy : *Element collection*, ktorý obsahuje všetky záznamy a *element record*, ktorý popisuje konkrétny záznam. Tak ako každá XML schéma, aj MARCXML obsahuje komplexné a jednoduché typy. Medzi komplexné typy patrí:

- *collectionType*
- *controlFieldType*
- *dataFieldType*
- *leaderFieldType*
- *recordFieldType*
- *subfielddatafieldType*

Komplexné typy obsahuju elementy, atribúty a text. Medzi jednoduché typy, ktoré obsahujú iba text patrí:

- *controlDataType* - string
- *controltagDataType* - string
- *idDataType* - ID
- *indicatorDataType* - string
- *leaderDataType* - string
- *recordTypeType* - string enumeration = *(Bibliographic, Authority, Holdings, Classification, Community)*
- *subfieldcodeDataType* - string
- *subfieldDataType* - string
- *tagDataType* - string

Už z názvov jednotlivých komplexných a jednoduchých typov vieme intuitívne namapovať MARC 21 schému bibliografického záznamu do formátu MARCXML. Štruktúra záznamu je nasledovná [20]:

- *element collection* - complexType = collectionType; attributes = id
  - *element record* - complexType = recordType; attributes = id, type
    - *element leader* - complexType = leaderFieldType; attributes = id, space
    - *element controlfield* - complexType = controlFieldType; attributes = id, tag(povinný), space
    - *element datafield* - complexType = dataFieldType; attributes = id, tag(povinný), ind1(povinný), ind2(povinný)
      - *element subfield* - complexType = subfielddataFieldType; attributes = id, code(povinný)

Na obrázku 4.2 môžeme vidieť ako tag 260 spadá pod 2xx pole v rámci pola premenných, ktoré hovorí o popisných údajoch dokumentu. Tento tag v MARCXML formáte je atribút elementu datafield, ktorý má pod sebou 3 subfield-y s atribútmi code, ktoré hovoria o kódoch podpolí. Indikátory tagu 260 sú prázdne.

```
▼<marc:datafield tag="260" ind1=" " ind2=" ">
   <marc:subfield code="a">New York, N.Y. :</marc:subfield>
   <marc:subfield code="b">Atlantic,</marc:subfield>
   <marc:subfield code="c">[1957?]</marc:subfield>
 </marc:datafield>
```

Obr. 4.2: Pole reprezentované MARCXML formou [20]

4.7 Štandard OAI-PMH

Hoci štandard MARC je jeden z najpoužívanejších vôbec, stále existujú rôzne reprezentácie bibliografických metadát. Pri komunikácií viacerých knižničných systémov medzi sebou môže dôjsť k situácií, že štandardy a formáty jednotlivých

systémov nevedia spolu dokonale komunikovať a môže prísť k dátovým stratám. Preto existuje štandard OAI-MPH, ktorý práve rieši problematiku zhlukovania veľkého množstva metadát s cieľom eliminovať dátové straty pri komunikácií knižničných systémov s centrálnym systémom a jeho databázou. Jedná sa o klasický server-client prístup. Metadáta prenášané z lokálných repozitárov cez centrálnu databázu až k užívateľovi sú reprezentované ako XML dokumenty posielané cez HTTP protokol. V podstate vie tento štandard fungovať s ľubovolným formátom, ktorý je možný zapísať do XML dokumentu [21]. Výsledný metadátový záznam je štrukturovaný do troch častí:

- *Header* : obsahuje identifikátor, časový údaj a názov setu repozitára
- *Metadata* : metadátový záznam dokumentu obsahuje jednoduché typy ako title, creator, description, publisher, date, type, identifier, type, language
- *About* : Doplnková časť obsahujúca napríklad autorské práva

OAI-PMH štandard pracuje so 6-timi request metódami:

- *Identify* - identifikuje repozitár
- *ListMetadataFormats* - vypíše podporované metadátové formáty ako napríklad (oai_dc, marc21, z39.50, ...).
- *ListSets* - vypíše štruktúru setov záznamov daného repozitára
- *ListRecords* - vypíše záznamy konkrétneho setu repozitára
- *GetRecord* - vypíše konkrétny jeden záznam
- *ListIdentifiers* - vypíše všetky hlavičky záznamov

Request metóda, ktorá by mala zobraziť všetky záznamy vo formáte marc21 by vyzerala takto: url?verb=ListRecords&metadataPrefix=marc21 [22, 23].

Kapitola 5

Analýza dát a použitie štatistických metód

5.1 Analýza dát a hľadanie súvislostí

Analýzou dát sa všeobecne nazýva proces manipulácie dát za cieľom zistenia nových užitočných údajov, ktorými vieme podporiť rozhodovanie sa. V knižničnom informačnom systéme sa analýza dát využíva na štatistické výstupy predajnosti, vypožičiavania dokumentov či informácií o čitateľoch. Knižniča tieto dáta spracuje a vie ich efektívne použiť napríklad na úrovni marketingu, pre zvýšenie záujmu čitateľov o výpožičky , ale aj na revíziu artefaktov v jednotlivých segmentoch knižnice, čím vedia docieliť znásobenie počtu kópií, prípadne vyradenie niektorých dokumentov.

Dátová analýza je samozrejme všeobecný pojem. Má niekoľko prístupov a prevedení. Tradičný prístup ku knižničným dátam nebral do úvahy analýzu a predikciu dát. Big data takýto prístup umožňuje. Umožňuje na základe analýzy záujmu

používateľov o tituly prinášať nové informácie a odporúčania pre používateľa. Predikcia umožňuje okrem iného knihovníkom spravovať knižničné dáta efektívne. Knižničné dáta sú teda typickým príkladom neštrukturovaných dát. Pri práci s knižničnými dátami dochádza k rôznym situáciám, ktoré sú v rámci analýzy limitujúce. Napríklad udržiavanie centrálneho dátového repozitára, v ktorom figurujú vzťahy v rámci knižničného katalógu. Okrem iného sa započítavajú aj dáta zachytené pri práci používateľa s knižničným systémom [40, 30, 45].

5.1.1 Data mining

Špecifickým pohľadom na dátovú analýzu je *data mining*. *Data mining* je prienik štatistiky a počítačovej vedy. Je to proces objavovania vzorov, vzťahov, kolerácií, vytvárania modelov na veľkom množstve predtým nepoznaných dát prevažne uložených v relačných databázach. *Data mining* používa metódy extrakcie, kombinácie, strojového učenia, štatistiky. Obsahuje teda rôzne techniky a prístupy napríklad aj s použitím *big data*:

- Asociácia – vytváranie vzťahov medzi viacerými objektami. Príkladom z praxe pre knižničný systém je vypožičanie učebnice slovenského jazyka s učebnicou anglického jazyka. Na základe týchto dvoch titulov, vieme určiť, že používateľ je napríklad študent, prípadne sa len vzdeláva v jazykoch. Knižničný systém mu na základe tejto asociácie vie odporučiť učebnicu nemeckého jazyka.

- Klasifikácia – vytváranie nejakých množín objektov, ktorých charakterizuje spoločná vlastnosť, na základe ktorej boli do danej množiny pridelený. Pri tejto technike vieme použiť rovnaký príklad, len algoritmicky inak poňatý. Klasifikácia teda rozširuje a využíva techniku asociácie.

- Clustering – vytváranie zhlukov na základe spoločných atribútov z veľkého

množstva dát.

- Rozhodujúci strom – slúži na kategorizáciu a predikovanie dát. Každé vetvenie vychádza z nejakej požiadavky, otázky, dát a podľa počtu výsledkov, odpovedí alebo typologie dát podľa určitých atribútov sa vytvára rovnaký počet vetiev. V nižších hĺbkach je kategorizácia viac presnejšia.

- Sekvenčné vzory – slúžia na identifikáciu trendov alebo objavovanie podobných udalostí. V knižniciach slúži aj na analýzu správania sa čitatela pri vykonávaní výpožičiek. Slúži na odporúčanie dát, na prieskum, ktoré dáta sú najviac vypožičiavané a ktorí zákazníci si ich vypožičiavajú [42, 36, 33].

5.1.1.1 Clustering

Clustering analýza znamená zhlukovanie dát z jedného súboru do zhlukov, ktoré v rámci jedného zhluku majú určité vzťahy rozličné od iných. Opis vzťahov stručne charakterizuje vzťahy medzi pôvodnými dátami a extrahuje skryté vzťahy. V kombinácií so zhlukovaním, práve jeden zhluk predstavuje jeden všeobecný vzťah. *Clustering* je považovaný za prístup bez dozoru. Nie je však sám o sebe algoritmom, ale iba všeobecným prístupom dátovej analýzy. Z *angl. Unsupervised* prístup bez dozoru, znamená, že algoritmus sa učí pracovať s dátami o ktorých predtým nemal žiadne informácie ani dopredu predurčené aké sú správne vzťahy alebo vzory. Konkrétnym algoritmom hĺbkovej analýzy dát je napríklad *canopy clustering*, ktorého pôvodcom je Andrew McCallum. Iným príkladom konkrétneho použitia clasteringu je K-means algoritmus, ktorého použitie pre knižničné dáta je vhodnou voľbou pre jeho rýchlosť. Celkovo pre knižničný systém je práve clustering analýza a opis vzťahov veľmi výhodná. Kombináciu zhlukovania a opisu vzťahov využíva aj ARTMAP framework. ARTMAP a jeho algoritmus je založený na tvorbe referenčných vektorov, vďaka ktorým je aplikovanie *data mining*-u schopné klasifikovať

vzory používateľských profilov do zhluku podobných vzorov, ktoré tvoria základ pre odporúčanie nových kníh. Existuje niekoľko metód použitia algoritmu zhlukovania:

- *Metódy partícií* - ako vstup prijímajú set partícií alebo dielov spolu s parametrom k, ktorý nesie v sebe informáciu o počte vstupných partícií. Metóda v sebe vykonáva proces takzvanej iteratívnej relokácie. Podstata tohto procesu je snaha premiestniť objekty z jednej partície do druhej za cieľom vylepšenia samotných partícií. Príkladom algoritmov využívajúcich tento systém sú CLARANS a K-means.

- *Metódy založené na hierarchii* - ako vstup prijímajú dáta, ktoré hierarchicky rozkladajú dvoma spôsobmi a to zhora nadol alebo zdola nahor v závislosti od nastavenia metódy. Kvalitu hierarchického rozkladu a následného zhlukovania sa vieme zdokonaliť dvoma spôsobmi:

  – analýzov vzťahov a väzieb po rozklade v každom zhluku

  – aplikovanie na hierarchicky rozložených zhlukoch takzvaný *Microclustering*, čiže ďalšie rozdelovanie zhlukov na základe podobných vlastností a aplikovanie rôznych iných typov metód. Je možné napríklad použiť aj predošlí spôsob iteratívnej relokácie.

  Príkladom algoritmov využívajúcich tento systém sú BIRCH, ROCK, chameleon.

- *Metódy pracujúce na báze hustoty* - ako vstup prijímajú dáta, na ktorých aplikujú zhlukovanie na základe hustoty. Takisto existujú dva prístupy tejto metódy:

  – zväčšovaním zhluku na princípe porovnávania hustôt jeho susedných zhlukov

– aplikovaním špecifickej funkcie na prácu s hustotou, ktorou sa ovplyvní štruktúra a veľkosť zhlukov

Príkladom algoritmov využívajúcich tento systém sú DBSCAN, DENSLUE.

- *Metódy pracujúce na báze mriežky* - ako vstup prijímajú dáta, ktoré štrukturálne usporadúvaju do mriežky pozostávajúcej z počtu buniek závislého od počtu dát. Po vytvorení mriežky aplikujú zhlukovanie, čiže spájajú bunky dát mriežky na základe určitých vlastností. Príkladom algoritmov využívajúcich tento systém sú STING, WAVE-CLUSTER.

- *Metódý pracujúce na báze porovnávania modelov* - ako vstup prijímajú dáta bez nejakej špecifickej formy. Podstata spočíva v hľadaní zhody a optimalizácie medzi objektamy dát a predvolenými matematickými modelmi, ktoré fungujú na princípe pravdepodobnosti, že objekty dát boli generované kombináciou základnej distribúcie. Príkladom algoritmov využívajúcich tento systém sú EM, SOM, COBWEB [37, 42, 36].

5.1.1.2 Algoritmus zhlukovania

Aplikácie algoritmu zhlukovania na knižničných dátach v rámci knižničného informačného systému má viacero využití. Pomocou algoritmu vieme analyzovať predajnosť titutolv a následne z výsledkov usudzovať, ktoré tituly nemajú veľa spoločných asociačných vlastností s inými titulmi, čím klesá ich možnosť byť odporúčanými a teda klesá predajnosť. Rovnako vieme analyzovať a začleňovať zákazníkov knižnice, buď podľa ich vytvoreného profilu v knižničnom systéme, kedy sa môže stať, že zákazník sa považuje za aktívneho, s veľkých targetovaním na tituly rôzneho typu, alebo je to zákazník, ktorého vypožíčané tituly nemajú veľa súvislostí, tým pádom je nepredvídateľný, čo je náročnejšie na targetovanie odporúčaných titulov. Práve analýza správania zákazníka podľa výpožičiek je kľúčová

pre knižnice.

5.1.1.3 Algoritmus sekvenčných vzorov

Sekvenčné vzory sú jedným s prístupov hĺbkovej dátovej analýzy. Ich podstatou je hľadanie trendov alebo udalostí a vytvárať z nich vzory. Konkrétnym príkladom algoritmu využívajúceho tento princíp je algoritmus AprioriAll. Pri použití na knižničné dáta, vieme aplikovať tento algoritmus na systém výpožičiek a ich dát uložených v databáze. Výpožička v databáze má okrem iných metadát aj identifikátor čitateľa, čas transakcie a samotné bibliografické údaje jednotlivých vypožičaných titulov. Keďže sa jedná o prístup pomocou sekvenčných vzorov, tak všetky výpožičky čitateľa vieme pokladať za jednu sekvenciu. Z toho vieme pomocou algoritmu určiť pri vopred stanovenej hraničnej hodnote výskytu, ktoré sekvencie sa vyskytujú najviac často. Výsledkom môže byť systém odporúčania, kedy knižničný systém z tejto dátovej analýzy vie vyhodnotiť, ktorý čitateľ si aké knihy vypožičiava, ako často a tým mu odporučiť knihy s rovnakou tematikou. Príklad konkrétnej aplikácie algoritmu je popísaný v dodatku.

5.1.1.4 Algoritmus korelačného odporúčania

Kolaboratívne odporúčanie je jedným z troch algoritmov, ktoré sa využívajú pri odporúčacích systémoch. Jeho podstatou je zbieranie veľkého množstva dát o užívateľoch, ich aktivitách a na základe toho porovnáva tieto dáta s ostatnými užívateľmi. Výsledkom môže byť histogram spoločných dát spomedzi jednotlivých užívateľov [24].

5.1.1.5 Algoritmus obsahového odporúčania

Obsahové odporúčanie narozdieľ od korelačného pracuje s konkretnými dáta jednotlivca a zamierava sa na obsah jeho aktivít. Zbiera dáta z výpožičkových proce-

sov čitateľa a tie dáta porovnáva s knižničnými databázami. Následne odporúča knižné tituly napríklad z rovnakej edície, alebo od rovnakého autora, alebo z rovnakej skupiny konspektu [24].

5.1.1.6 Algoritmus hybridného odporúčania

Z názvu je už zrejmé, že hybridné odporúčanie je spojené korelačné a obsahové odporúčanie dokopy. Hybridné odporúčanie je možné použiť napríklad pri veľmi malom a nereprezentatívnom množstve dát. Existuje viacero prístupov ako aplikovať hybridné odporúčanie, napríklad aplikovaním pravidiel jedného princípu do druhého, alebo separátne aplikovanie a následne spojenie dát, prípadne vytvorenie spoločného algoritmu [24].

5.2 Štatistické metódy

V predošlej kapitole sme analyzovali všetky typy knižničných dát a následne je potrebné na ne aplikovať štatistické metódy použitím štatistických nástrojov respektíve softvérov. Použitím rôznych typov štatistických metód vieme analyzovať nové vzťahy medzi existujúcimi dátami a tým vytvárať nové dáta prospešné pre chod a celkový obraz knižničného informačného systému.Štatistické metódy vieme primárne zadeliť do 4 kategórií podĺa prístupu k dátam. Jedná sa o parametrické inferenčné metódy, neparametrické inferenčné metódy, prediktívne korelačné metódy, prediktívne regresné metódy.

5.2.1 Inferenčné metódy

Parametrické inferenčné metódy sú nasledovné :

- t-test

- ANOVA
- MANOVA
- ANCOVA
- Z-test
- porovnávací test

Z vyššie uvedených je veľmi používaný t-test. T-test funguje na princípe overovania či určitý dopredu určený predpoklad je pravdivý na danej vzorke dát [32]. Neparametrické inferenčné metódy sú nasledovné:

- chi-kvadrát test
- Wilcoxonov test
- Mann-Whitneyov U test
- Kruskal-Wallisov test
- Kolmogorov-Smirnovov test
- Kendallov W test
- Friedmanov test
- Binomický test

Zo spomenutých neparametrických metód je asi najznámejšia chi-kvadrát metóda, ktorá v princípe zisťuje, či je nejaký vzťah medzi dvoma dátovými jednotkami z rôznych kategórií [32]. Primárnym rozdielom medzi parametrickými a neparametrickými inferenčnými metódami je použitie predpokladu. V princípe ak máme naše dáta nejakým spôsobom rozumne rozdelené podľa rôznych parametrov vieme použiť parametrické inferenčné metódy. Naopak ak nemáme žiadne bližšie infor-

mácie k naším dátam použijeme neparametrické inferenčné metódy. Prvý spôsob už z opisu bude mať vždy na rovnakých dátach vyššiu výpovednú hodnotu ako ten druhý [25].

5.2.2 Korelacné metódy

Medzi prediktívne korelačné metódy patria:

- Korelácia
- Pearosonova korelácia
- Spearmanova korelácia
- korelácia poradia

Medzi prediktívne regresné metódy patria:

- Regresia
- Lineárna regresia
- Logistická regresia
- Viacnásobná regresia
- Hierarchická regresia
- Coxova regresia

Rozdieľ medzi prediktívnymi korelačnými štatistickými metódami a prediktívnymi regresnými štatistickými metódami je v prístupe k dátam. Korelácia určuje vzťah medzi 2 dátovými jednotkami, ktoré nejak bližšie nedefinuje. Naopak regresia určuje ako je jedna nezávislá dátová jednotka numericky prepojená so závislou dátovou jednotkou [26].

5.3 Štatistické nástroje

Keďže štatistické metódy sú známe a je ich veľké množstvo, štatistické nástroje ponúkajú prostredie pre dátovú analýzu použitím už spomínaných štatistických metód. Viaceré štatistické nástroje fungujú ako celkový softvér, kde ponúkajú vývojové prostredie, samotný programovací jazyk, aj výstupné formáty v podobe tabuliek alebo grafov. Medzi najznámejšie štatistické nástroje patria R, SaS, SPSS, Stata, Matlab a ďalšie. SaS pokrýva celé spektrum dátovej analýzy od spracovania dát, cez ich samotnú dátovú analýzu prostredníctvom štatistických metód ako t-testovanie a rôzne druhy regresie a korelácie, až po zobrazenie výsledných sústav. SPSS je jeden z najstarších softvérov na dátovú analýzu. Podobne ako SaS aj on pokrýva celú škálu štatistických metód od inferenčných po prediktívne.

5.3.1 R

Jeden z najpoužívanejších softvérov na štatistické účely je R. Okrem toho, že je to aj samostatný programovací jazyk, R ponúka prostredie na štatistickú dátovú analýzu a rôzne výstupné formáty. Softvér R má viacero balíkov. Základný balík ponúka niekoľko metód dátovej analýzy, napríklad Pearsonovu korelačnú metódu. Jeden z ďalších balíkov je *atsa* balík, ktorý primárne slúži na dátovú analýzu časových údajov, ale je možné ho použiť aj na lineárnu regresiu. V rámci štúdie o aplikovaní štatistických metód softvéru R na knižničné dáta *The College of New Jersey Library* boli vykonávané 2 analýzy s použitím štatistických metód ako lineárna regresia a Pearsonovu koreláciu [41].

5.3.1.1 Analýza použitím lineárnej regresie

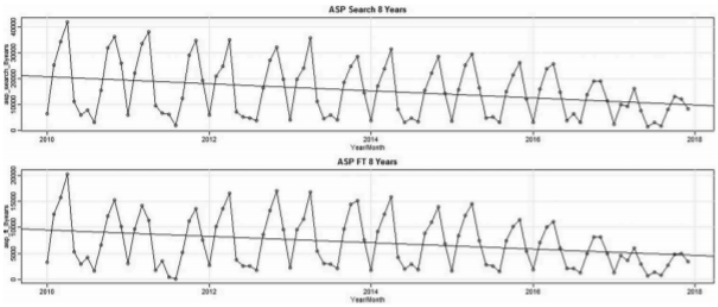
Lineárna regresia je metóda, ktorá už podľa svojho názvu pracuje s lineárnou funkciou Y = a + bX.

	A	B	C
1	Month_Year	SEARCH	FT
2	Jan-10	6423	3280
3	Feb-10	25389	12544
4	Mar-10	34256	15768
5	Apr-10	42018	20237
6	May-10	11248	5342
7	Jun-10	5939	3002
8	Jul-10	7860	4218
9	Aug-10	3140	1618
10	Sep-10	15605	6599
11	Oct-10	32018	12189
12	Nov-10	36204	15259
13	Dec-10	25927	10222
14	Jan-11	5961	3017

Obr. 5.1: Dátový set o počte vyhľadávaných dát a dokumentov za určité obdobie [41]

Použitím lineárnej regresie na časové obdobie siedmich rokov (mesačne) a počte vyhľadávacích dát vieme zistiť, že medzi týmito dvoma typmi dát je vzájomný vzťah, ktorý naznačuje, že od začiatku sa postupne pribúdaním rokov zmenšil počet vyhľadávaní o hodnotu 1367 ročne. Tento údaj môže napovedať rôzne predikcie do budúcnosti a prípadnom úbytku čitateľskej základne alebo naopak o väčšej efektivite vyhľadávania. K výsledným dátam by bolo vhodné napríklad doplniť dáta týkajúce sa kľúčových slov a úspešnosti vyhľadávania. Následne by sme vedeli určiť či tento údaj ma pozitívny alebo negatívny charakter. V štúdii následne použili rovnaký princíp aj na sťahovanie dokumentov. Tam bol ročný pokles o hodnotu

603. Keď tento údaj priložíme k tomu predošlému, vidíme, že to potvrdzuje ten negatívny charakter a síce pravdepodobný úbytok čitateľskej základne alebo len zníženie jej aktivity. Graficky to je znázornené nižšie [41].



Obr. 5.2: Grafické znázornenie ročného úbytku vyhľadávania a sťahovania dokumentov [41]

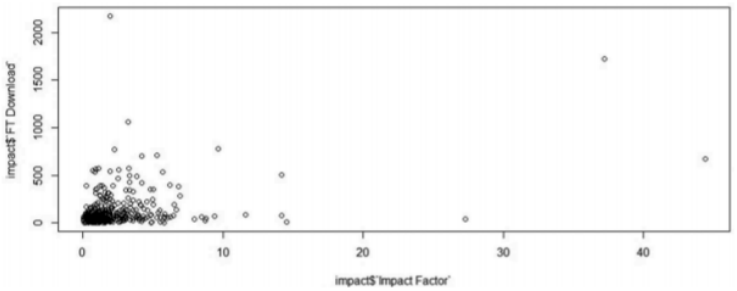
5.3.1.2 Analýza použitím Pearsonovej korelácie

Pearsonova korelácia použitím koeficientu r rieši vzťahy medzi dvoma závislosťami. V rámci štúdie bol cieľ zistiť aký veľký je vzťah medzi sťahovaním dokumentov a impactom týchto dokumentov. Aký veľký vzťah to je nám pomáha určiť už vyššie spomenutá Pearsonov korelačný koeficient r, ktorý definuje nasledovné vzťahy na množine $-1 < r < 1$, kde hodnota 0 znamená žiadnu koreláciu, 1 dokonale pozitívnu a -1 dokonale negatívnu koreláciu. Ostatný interval je vysvetlený nasledovne:

- od -0.99 po -0.6 = silne negatívny vzťah
- od -0.59 po -0.3 = mierne negatívny vzťah
- od -0.29 po 0.29 = slabý vzťah
- od 0.3 po 0.59 = mierne pozitívny vzťah
- od 0.6 po 0.99 = silne pozitívny vzťah

Analýza sa robila na vzorke 800 elektronických žurnálov, z ktorých malo 390 aspoň

nejaký impact factor. Impact factor je hodnota počtú citácií daného článku za daný rok. Aplikovaním Paersonovej koreláce na týchto dátach sme dostali hodnotu konštanty r na úrovni 0.36, čo vieme nazvať ako mierne pozitívny vzťah [41].



Obr. 5.3: Grafické znázornenie Impact factoru na počet sťahovaní dokumentov [41]

Kapitola 6

Zhodnotenie analýzy

V rámci analýzy tohto bakalárskeho projektu som sa snažil pokryť celú problematiku práce s knižničnými dátami a aplikácie štatistických metód na nich.

Na začiatku som všeobecne analyzoval aké existujú pamäťové inštitúcie, čím som sa chcel dostať ku knižniciam a knižničným systémom. V rámci tejto kapitoli som sa snažil definovať podstatu informačného systému v knižniciach, konfrontovať to s dobou pred digitalizáciou a po nej. Odrazovým bodom do ďalšej analýzy bolo načrtnutie problematiky vyhľadávania dát v knižničnom systéme, keďže v konečnom dôsledku práve správne aplikovanie štatistických metód na knižničné dáta má spôsobiť zvýšenú efektivitu pre zákaznika knižnice prostredníctvom odporúčania jednotlivých titulov na základe jeho doterajšej činnosti.

V ďalšej kapitole som už analyzoval samotné knižničné dáta. Keďže informačný systém pracuje s nespočetným množstvom dát, rozdelil som ich do 3 kategórií na evidenčné, výpožičkové a systémové dáta. Evidenčné dáta ako najpodstatnejšiu kategóriu dát som sa snažil bližšie analyzovať pomocou štandardov a noriem, ktoré definujú akým spôsobom sa s danými dátami manipuluje. Zároveň som analyzoval

samotné syntaxové reprezentácie MARC 21 štandardu, keďže je potrebné vedieť ako pracovať s danými dátami. Podstatným typom dát boli aj systémové dáta, ktoré hovoria o fungovaní samotného informačného systému knižnice ako napríklad počet návštev, počet vyhľadávaní, aké sú to vyhľadávania , odkiaľ sú návštevy a podobne. Tieto všetky informácie vieme získať zo systémových logov. Kapitolu o analýze dát som zakončil analýzom a zamyslením sa nad novým prístupom k dátam, čím som mierne naznačil aj návrh systému a jeho určité funkcionality.

V nasledujúcej kapitole som definoval rôzne prístupy k analýze knižničných dát. Zameral som sa na data mining pohľad a jeho rôzne techniky z pomedzi, ktorých som bližšie analyzoval *Clustering* a jeho algoritmy. Táto prvá časť dátovej analýzy riešila zber dát a jeho správne zatriedenie. V druhej časti som už analyzoval štatistické metódy a nástroje, ktoré sa použijú na vopred pripravených dátach. V rámci štatistických metód som pokryl všetky 4 typy parametrických a preditktívnych metód ako aj metód odporúčania. Ako príklad štatistického nástroja som si zvolil R. Podľa štúdií, ktoré som k nemu našiel som analyzoval použitie 2 štatistických metód pomocou tohto nástroja na konkrétnych príkladoch knižničných dát.

Kapitola 7

Opis riešenia

Zo zadania je mojou úlohou na základe výsledkov analýzy navrhnúť informačný systém, ktorý bude nadstavbou KIS a bude knižniciam poskytovať variabilitu pri tvorbe výstupných zostáv, štatistík a rôznych pohľadov na dáta z ich informačných systémov so zameraním na odhaľovanie nových poznatkov. Navrhovaný IS bude webová aplikácia určená pre knihovníkov a knihovníčky, respektíve celkovú administratívu knižnice. Používateľ aplikácie si bude môcť upravovať vstupné dáta pre čo najšpecifickejšiu analýzu. Po naimportovaní vstupných dát systém automaticky spustí validáciu dát. V rámci validácie budú kontrolované údaje v konkrétnych záznamoch na základe očakávaných vlastností, ktoré by daný údaj mal spĺňať. Po validácií systém dáta spracuje podľa parametru o výbere analýzy, čiže namapuje dáta na dátový model konkrétnej analýzy a uloží ich do relačnej databázy. Po spracovaní bude systém pripravený analyzovať dáta. Používateľ bude mať možnosť si upravíť vstupné údaje. Pred spustením algoritmov potrebných na vykonanie analýz, systém podľa vstupných údajov od používateľa vyselektuje konkrétne záznamy, ktoré spĺňajú špecifikáciu, z relačnej databázy. Po vykonaní analýzy systém vypíše alebo zobrazí výstupne zostavy formou rôznych typov grafov. Počas všet-

kých krokov bude možné sledovať systémové logy, ktoré budú popisovať dôležité kroky systému a v prípade chyby, bude možné na základe logov zistiť kde nastal problém. Používateľ môže analyzovanie dát opakovať ľubovoľný počet krát. Ak bude chcieť zopakovať rovnakú analýzu na rovnakej vzorke dát systém si ponechá dáta v Cache pamäti.

7.1 Špecifikácia požiadaviek

Funkcionalita knižničného informačného systému nie je veľmi obšírna. Dá sa zatriediť do niekoľkých funkcionálnych blokov. Prvým je samozrejme prihlasovací blok. Nasleduje blok importovania vstupných dát, blok analýzy dát a blok dátových výstupov. Špecifikáciu požiadaviek pokrývajú funkčné a nefunkčné požiadavky.

7.1.1 Funkčné požiadavky

- 1. Aplikácia musí byť schopná importovať csv a xml súbory
- 2. Aplikacia musí poskytovať prihlasovanie pre administratívu knižnice
- 3. Aplikácia musí byť schopná validácie, spracovania a agregácie knižných dát v Cache pamäti
- 4. Aplikácia musí byť schopná komunikovať s relačnou databázou
- 5. Aplikácia musí byť schopná ukladať dáta na analýzu v relačnej databáze
- 5. Aplikácia musí umožniť špecifikovať vstupné dáta užívateľom na spustenie analýzy
- 6. Aplikácia musí byť schopná analyzovať vývoj aktivít používateľskej základne knižnice

- 7. Aplikácia musí byť schopná analyzovať transakcie KIS na tituloch knižničného fondu
- 8. Aplikácia musí logovať každú manipuláciu so systémom a dané logy zobrazovať
- 9. Aplikácia musí zobrazovať vstupné a výstupné dáta analýz v prehľadnej forme pre užívateľa

7.1.2 Nefunkčné požiadavky

- 1. Aplikácia musí byť schopná vykonať analýzy systémov v čo najkratšom čase
- 2. Aplikácia musí byť navrhnutá intuitívne a jednoducho
- 3. Aplikacia musí byť spoľahlivá v kontexte správnosti výstupov analýz

7.2 Návrhy analýz na knižničných dátach

V predošlých kapitolách som zhrnul s akým typom dát sa knižničný systém dostáva do kontaktu, s akými manipuluje prípadne na aký účel ich eviduje a používa. Avšak na dosiahnutie čo najväčšej efektivity využitia dát je potrebné hľadať nové vzťahy a z nich získavať nové dáta, ktoré môžu napomôcť fungovaniu KIS. Na základe nových dát vieme napríklad zlepšiť systém odporúčania knižných titulov pre konkrétných používateľov na báze ich predošlých výpožičiek alebo osobného profilu. Rovnako to vie napomôcť knižnici ako takej napríklad pri zisťovaní súčastných trendov, odrazov doby, potrieb užívateľov ale aj predikcie do budúcnosti, aby knižnica vedela poskytovať literatúru, o ktorú bude čo najväčší záujem a tým zvýši aj svoj ekonomický potenciál.

7.2.1 Analýza vývoja aktivít používateľskej základne knižnice

Prvým pohľadom na dáta v knižniciach, bude analýza vývoja používateľskej základne knižnice. Každým rokom pribúdajú nový používatelia, väčšina ostáva a niektorí prestanú byť aktívnymi. Cieľom tejto analýzy je zistiť ako sa vyvýja aktivita skupiny čitateľov vzhľadom na ich vek a pohlavie v konkrétnych okresoch, prípadne mestských častiach Českých Budejovíc po určitú dobu.

7.2.1.1 Dáta

Na danú analýzu budeme využívať dáta z transakcií knižnice. Z daných dát nás budú zaujímať hlavne používateľské informácie. Konkrétne sa jedná o identifikátor používateľa, ďalej to je jeho vek, pohlavie a poštové smerovacie číslo, ktoré nám bude určovať oblasť v ktorej má trvalé bydlisko. Všetko to bude v závislosti od času, čiže sa zameriame na časové stopy aktivít používateľov, keďže chceme zistiť vývoj aktívnéj časti používateľov.

7.2.1.2 Spracovanie dát

Zo všetkých údajov o transakciách vyselektujeme tie, ktoré budú obsahovať všetkých používateľov s ich vekom, poštovým smerovacím čislom, pohlavím a časovou stopou kedy vykonali operáciu výpožičky. Dané dáta uložíme do samostatného modelu, čiže tabuľky. Všetky tieto typy dát si budeme môcť pred analýzou špecifikovať podľa seba, napríklad si vyberieme vekovú hranicu, pohlavie, prípadne poštové smerovacie čislo alebo časový interval. Samozrejme nemusíme si zvoliť nič a v takom prípade systém bude analyzovať dáta na celej množine..

7.2.1.3 Použitie metódy

Na finálnych predspracovaných dátach aplikujeme kolaboratívne štatistické metódy polynomiálnej a lineárnej regresie. Premenná, ktorá bude nezávislá, bude vždy časový údaj. Okrem výsledkov hlavnej analýzy budeme analyzovať aj každý jeden vstupný údaj, na ktorý aplikujeme histogram vzhľadom na počet aktivít používateľov.

7.2.1.4 Výstup

Výstupom analýzy bude graf polynomiálnej funkcie rovnako s histogramami jednotlivých údajov o počte užívateľov v okresoch, počte užívateľov v daných vekových skupinách, počte užívateľov na základe pohlavia.

7.2.2 Analýza transakcií na dostupných tituloch knižničného fondu

Druhým pohľadom na dáta o knižniciach je analýza vzťahu typu operácie konkrétnej transakcie používateľa na katalógový údaj o knižnom titule vzhľadom na časový interval. Cieľov tejto analýzy môže byť niekoľko, napríklad ktorí autori boli najviac vypožičiavaní za posledný rok, aké vydavateľstvo dominovalo, prípadne o aké tituly z pohľadu skupiny konspektu bolo najviac rezervovaných posledný mesiac.

7.2.2.1 Dáta

Pri tejto analýze budeme okrem dát z transakcií knižnice potrebovať aj celkový katalógový zoznam danej knižnice, čiže všetky dostupné tituly. Konkrétne sa teda jedná o identifikátor transakcie, typ transakcie, časová stopa transakcie, dĺžka vypožičania (tento údaj bude dostupný iba v prípade zvolenia typu operácie výpo-

žička), identifikátor do katalógového zoznamu, odkiaľ budeme pracovať s názvom autora (tag 100, kód podpoľa ä"), skupinou konspektu (tag 072, kód podpoľa "x"), názov vydavateľstva (tag 260, kód podpoľa "b"). Vhodnou alternatívou prístupu k dátam by bolo OAI harvestovanie

7.2.2.2 Spracovanie dát

Podobne ako pri prvom type analýzy aj tu najprv vyselektujeme z celého datasetu iba dáta potrebné pre túto analýzu a teda identifikátor transakcie, typ transakcie, časovú stopu transakcie, dĺžku vypožičania, názov autora, skupinu konspektu, názov vydavateľstva. Pred spustením analýzy budeme mať taktiež možnosť si upraviť vstupné dáta podľa seba. Na výber bude typ transakcie, interval časovej stopy, ak bude typ transakcie výpožička, bude možné zvoliť aj interval dĺžky vypožičania, taktiež bude na výber zvoliť množiny skupín konspektu. Dané vstupné hodnoty potom aplikujeme do finálneho selektu z tabuľky.

7.2.2.3 Použitie metódy

Na finálnych dátach rovnako použijeme kolaboratívnu štatistickú metódu regresie, čiže hľadanie vzťahu medzi vstupnými premennými. Premennými budú časový údaj, prípadne jeho intervaly a počet vyhovujúcich záznamov podľa finálneho selektu. Okrem hlavných analýz budeme vytvárať aj pomocné histogramy počtu transakcií podľa skupiny konspektu, podĺa typu transakcie, podľa vydavateľstva a podľa autora ako aj dĺžky vypožičania titulu.

7.2.2.4 Výstup

Výstupom analýzy budú teda pomocné histogramy spolu s hlavným grafom, ktorý bude reprezentovať analýzu na základe našich zvolených vstupných hodnôt.

7.2.3 Analýza knižničného fondu na vekovej štruktúre čitateľov

Analýza vzťahu vekovej štruktúry používateľov s autormi knižných titulov. Každá knižnica, či už väčšia alebo menšia má rozmanitú štruktúru užívateľov podľa rôznych kritérií, napríklad podľa veku. Do knižníc chodí celá škála ľudí, od malých školopovinných detí, cez ľudí v produktívnom veku až po seniorov na dôchodku. Každá veková skupina má samozrejme rozlišnú početnosť, z čoho vie knižnica zistiť aká veková skupina je dominantná. V tejto analýze sa pozrieme na to aká skupina konspektu prevláda v daných vekových skupinách.

7.2.3.1 Dáta

Vstupnými dátami v tejto analýze sú dáta z transakcií knižnice. Údaj

$$cbvk_u s_c at * 0591200$$

je identifikátorom titulu v zozname titulov danej knižnice. Identifikátor je číselný údaj za znakom *. Keďže transakcia obsahuje niekoľko operácií, pre nás bude podstatná operácia výpožičky. Okrem identifikátora transakcie, potrebným pre analýzu z dát o transakciách bude vek používateľa v čase výpožičky. Zo zoznamu titulov podľa identifikátora nájdeme konkrétny záznam a pod tagom 072 a kódom podpoľa "x"nájdeme skupinu konspektu, do ktorej daný titul patrí. Kód podpoľa 9 v tagu 072 obsahuje číselné vyjadrenie konspektu.

7.2.3.2 Spracovanie dát

Keď vieme s ktorými dátami budeme pracovať tak si ich pripravíme na analýzu, a síce vytvoríme zoznam transakcií, ktoré majú typ operácie výpožička. Zo vstupu od používateľa bude možné zvoliť vekový interval čitateľa, ako aj rôzny počet

skupín konspektu.

7.2.3.3 Použitie metódy

Na agregovaných dátach následne použijeme korelačnú štatistickú metódu regresie. Nezávislými hodnotami budú skupiny konspektu a vekové intervaly čitateľov.

7.2.3.4 Výstup

Výstupom tejto analýzy budú dva hlavné grafy ako výstupy korelačných regresií spolu s pomocnými histogramami počtu výpožičiek pre skupinu konspektu a pre vekové intervaly čitateľov.

7.3 Architektúra

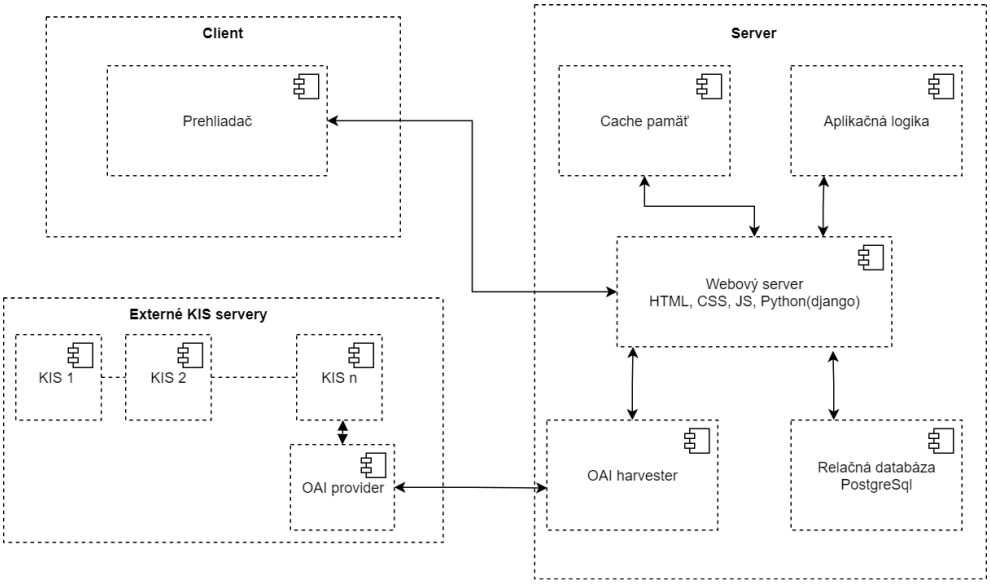
Jedná sa o server-client webovú aplikáciu, ktorej architektúra bude založená na princípe MDA *(anlg. Model driven architecture)*, čiže sa oddelí prezenčná vrstva od aplikačnej a databázovej. Celková architektúra je znázornená na diagrame.

7.3.1 Prezenčná vrstva

Keďže to bude webová aplikácia, prezenčná vrstva aplikácie bude webová stránka, prostredníctvom ktorej bude možné vykonávať jednotlivé funkcionality, sledovať výpisy systémových logov a zobrazovať výstupy analýz.

7.3.2 Aplikačná vrstva

Aplikačná vrstva bude obsahovať všetky funkcionality funkčných požiadaviek aplikácie. Aplikačná vrstva bude komunikovať jednak s prezenčnou vrstvou a to spôsobom, že bude dostávať vstupy nad ktorými bude vykonávať konkrétne funkcionality



Obr. 7.1: Architektúra webovej aplikácie

a ich výsledok vo forme výstupov pošle prezenčnej vrstve. Komunikácia bude aj s databázovou vrstvou nakoľko systém vo vlastných dátových typoch bude ponechávať iba aktuálne dáta používané v rámci vykonávania funkcionality a zvyšné dáta bude ukladať prípadne si ich vypytovať z relačnej databázy na ktorú bude mať prichystané dátové typy, ktoré budú namapované na dátový model databázy.

7.3.3 Databázová vrstva

Databázová vrstva aplikácie zabezpečuje komunikáciu medzi relačnou databázou a aplikačnou vrstvou aplikácie v rámci ktorej zabezpečuje aj správne mapovanie údajov.

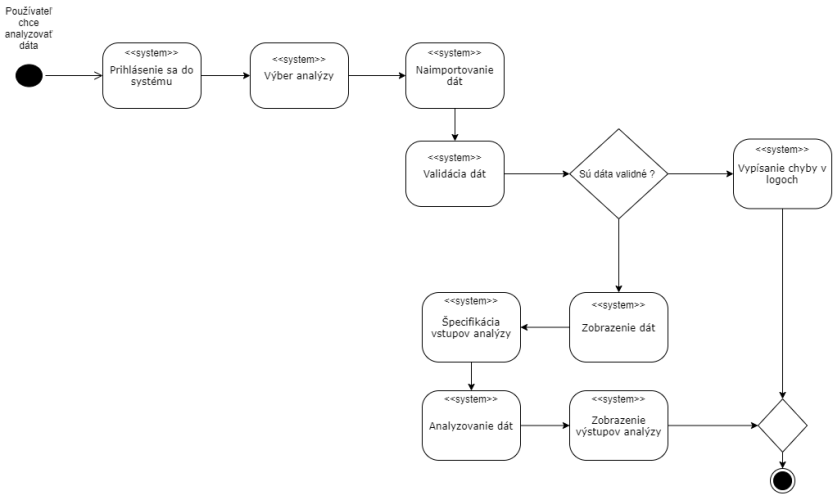
7.4 Diagramy aplikácie

Prípady použitia úzko nadväzujú na funkčné požiadavky aplikácie. Základným prípadom použitia je prihlásenie užívateľa do systému a všetky nasledujúce prípady

použitia už predpokladajú s úspešným prihlásením sa. Každý prípad použitia je opísaný aj UML *use-case* diagramom pre jeho príslušný funkcionálny blok. Na opis celkovej funkcionality webovej aplikácie som použil activity diagram.

7.4.1 Aktivity diagramy

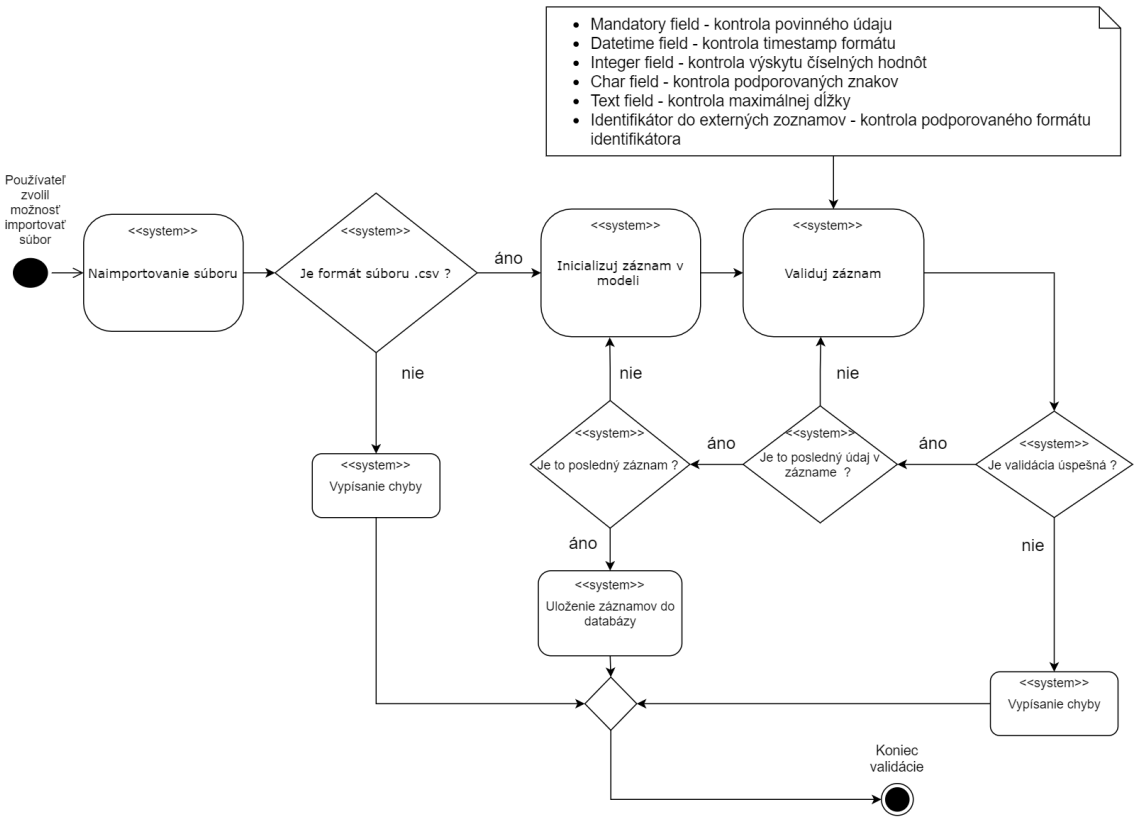
Hlavný aktivity diagram všeobecne popisuje *flow* funkcionalít ako idú chronologicky za sebou pri hociktorom type analýzy dát.



Obr. 7.2: Hlavný activity diagram

Aktivity diagram validácie dát podrobnejšie prechádza jednotlivými krokmi od importovania dát z externého .csv súboru až po stav, že dáta sú pripravené na vypísanie a agregovanie. Podstatou validácie je parsovanie dát z .csv súboru a kontrola, či na správnej pozícií sú správne dáta v správnej forme.

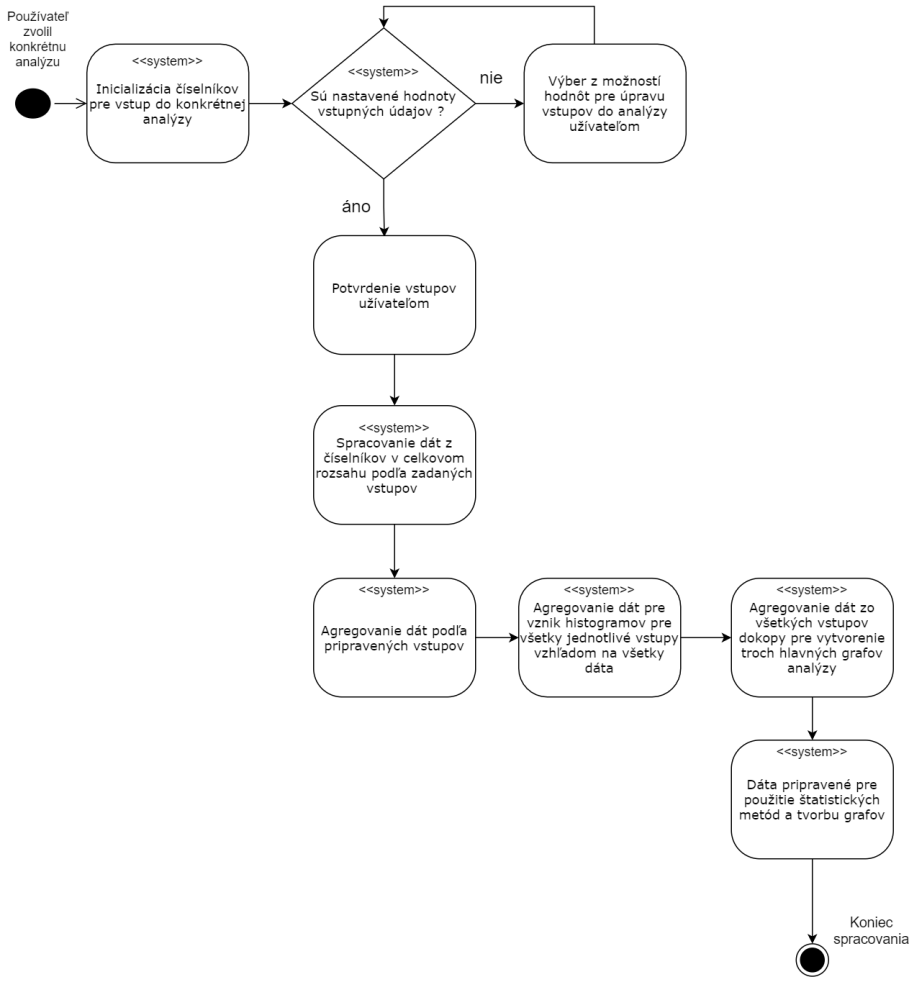
Po validácií si môžeme zvoliť vstupné zostavy do analýzy, ktoré budeme agregovať. Tieto ponúkané vstupy najprv inicializujeme z tabuliek číselníkov. Po zvolení vstupných hodnôt dané vstupy spracujeme, čiže vyselektujeme potrebné údaje z vedľajších tabuliek v závislosti od hlavnej tabuľky analýzy. Následne dané dáta



Obr. 7.3: Activity diagram validácie dát

agregujeme všetky dokopy, tak aby spĺňali každý vstupný údaj. Okrem toho agregujeme dáta aj jednotlivo pre každý jeden typ vstupného údaju v závislosti od hlavnej tabuľky. Všetky agregácie majú za cieľ vytvoriť dvojice číselných hodnôt, kde zväčša jedna hodnota je podľa typu vstupného údaju, tá bude nezávislá a závislá hodnota je počet záznamov, ktoré vyhovujú daným vstupom. Na takto pripravených dátach budeme aplikovať štatistické metódy.

Jednou z analýz je analýza vývoja aktivity čitateľskej základne knižnice podľa rôznych faktorov spojených s používateľom. Diagram obsahuje všetky kroky konkrétnejšie popísane vzhľadom na všeobecný aktivity diagram [Obr. 7.5]

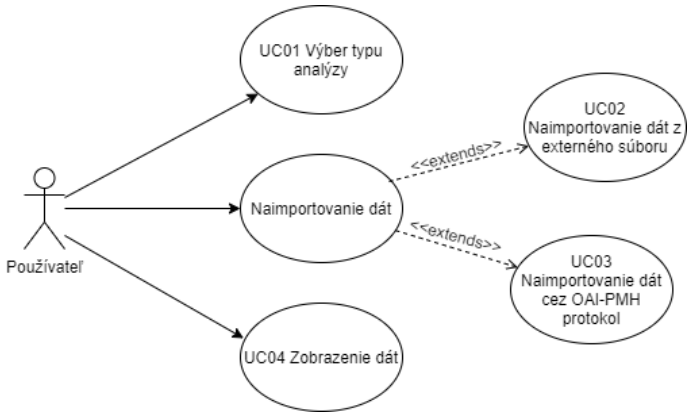


Obr. 7.4: Activity diagram spracovania dát

Obr. 7.5: Activity diagram 1. analýzy

7.4.2 Prípady použitia

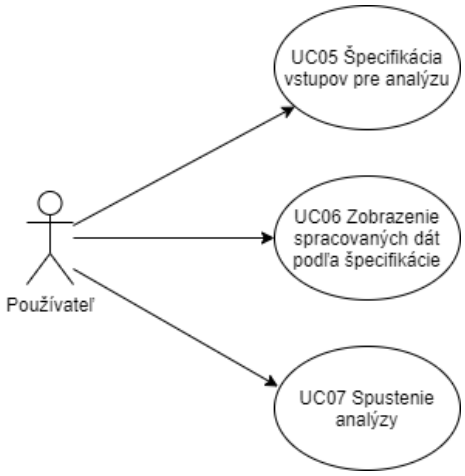
- 1. UC01 - Výber typu analýzy. Na začiatku si zvolíme, ktorú z analýz chceme použiť na vstupných dátach.
- 2. UC02 - Naimportovanie dát z externého súboru. Načítame csv alebo txt súbor na server.
- 3. UC03 - Naimportovanie dát cez OAI-PMH protokol. Prostredníctvom REST API stiahneme vstupné dáta pomocou OAI-PMH requestu (Voliteľná možnosť).
- 4. UC04 - Zobrazenie dát. Po systémovej validácií dát, systém dané dáta zobrazí.



Obr. 7.6: Use-case diagram bloku importovania dát

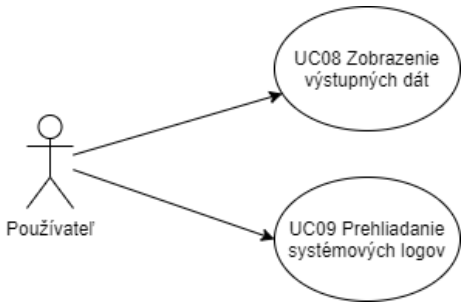
- 5. UC05 - Špecifikácia vstupov pre analýzu. Po úspešnej validácií vstupných dát špecifikujeme vstupy pre konkrétnu analýzu. Napríklad určíme časový interval, na ktorom budeme robiť analýzu. Špecifikácia vstupov je pre jednotlivé analýzy odlišná.
- 6. UC06 - Zobrazenie spracovaných dát podľa špecifikácie. Po špecifikácií vstupov, dané dáta spracujeme a zobrazíme.

- 7. UC07 - Spustenie analýzy. Spustíme algoritmy analýzy knižničných dát.



Obr. 7.7: Use-case diagram bloku analýzy dát

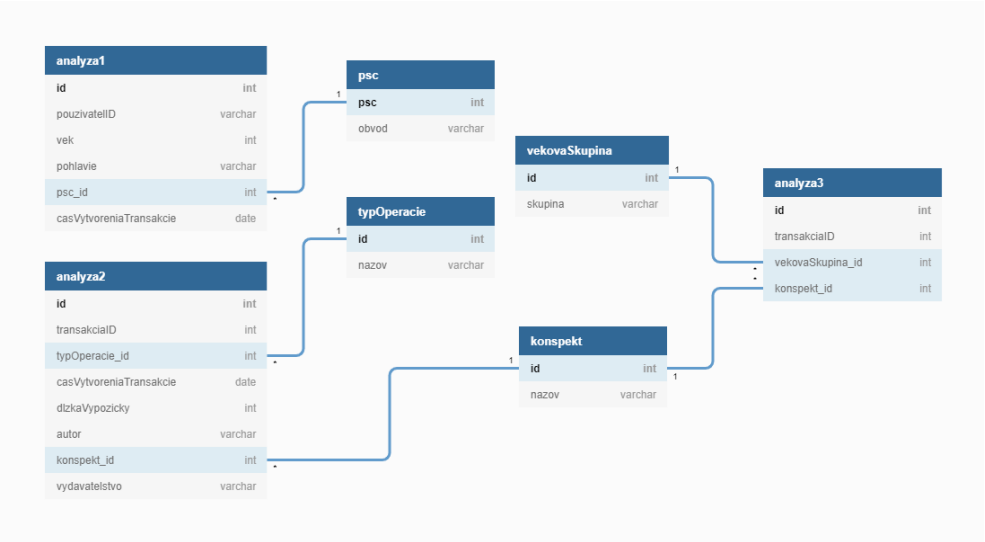
- 8. UC08 - Zobrazenie výstupných dát. Po úspešnej analýze zobrazíme výstupné sústavy.
- 9. UC09 - Prehliadanie systémových logov. Okrem výstupných sústav, si môžeme priebežne ale aj na konci pozrieť celkový priebeh analýzy v systémových logoch.



Obr. 7.8: Use-case diagram bloku dátových výstupov

7.4.3 Dátový model

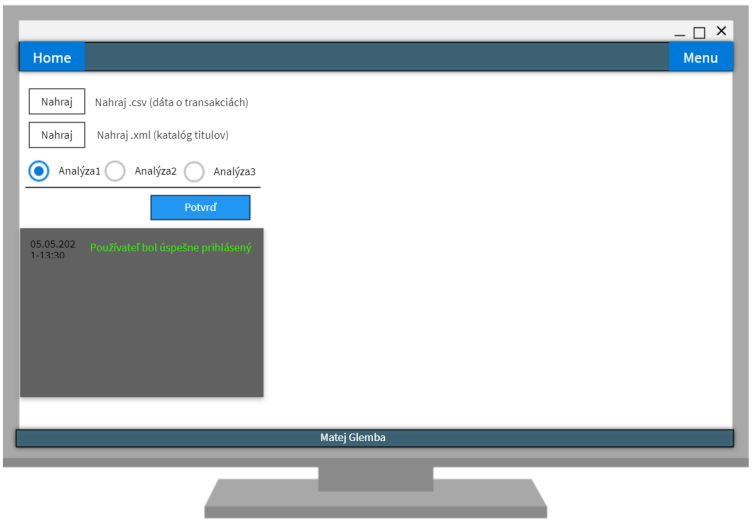
Dátový model aplikácie predstavuje všetky potrebné dátové entity používané na komunikáciu s relačnou databázou. Jedná sa o fyzický dátový model, na ktorom sú zobrazené vzťahy jednotlivých entít spolu s ich atribútmi, pri ktorých je znázornený aj podporovaný dátový typ. Podstatou aplikácie sú tri hlavné tabuľky pre každú z troch analýz. Okrem nich dátový model aplikácie obsahuje aj 4 číselníky, ktoré sú spojené s hlavnými tabuľkami vzťahom One-To-Many.



Obr. 7.9: Fyzický model

7.4.4 Mockupy

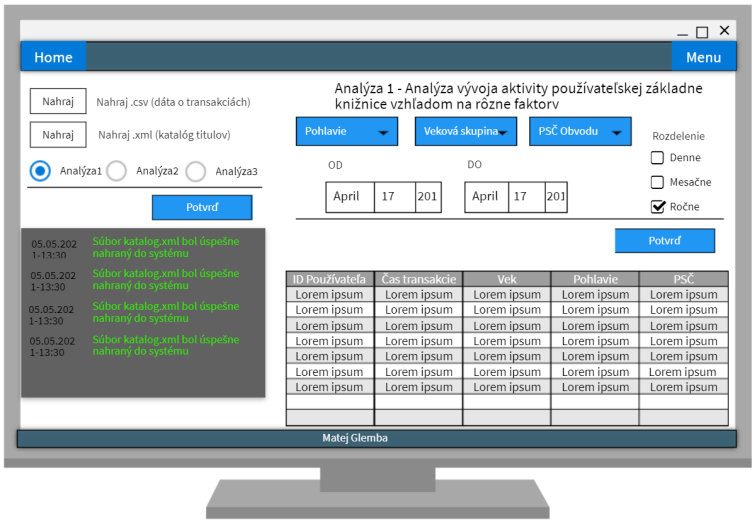
Mockupy znázorňujú návrh zobrazenia prezenčnej vrstvy webovej aplikácie a rozloženie potrebných elementov webovej aplikácie tak aby sa dodržala 3. nefunkčná požiadavka, ktorá hovorí o jednoduchosti a intuitivite webovej aplikácie. Prvý mockup je súčasťou prvej fázy analýzy. Po výbere konkrétneho typu analýzy si vyberáme možnosť importu dát. Na výber máme dva typy súborov. Csv súbor obsahuje všetky záznamy o transakciách knižnice, zatiaľ čo xml súbor je katalóg titulov danej knižnice. Na obrázku 7.10 vidíme, že v ľavej spodnej časti je priestor na systémové logy. Je to vhodná funkcionalita najmä počas importovania a validovania dát, kedy systém oznamuje používateľovi aktuálny stav.



Obr. 7.10: Import vstupných dát

Ak validácia prebehla úspešne, dáta zobrazíme v prehliadači. Následne špecifikujeme vstupné hodnoty pre algoritmy analýzy dát, tak aby sme nemuseli robiť analýzu na celkovom množstve vstupných dát, ale na určitej špecifickej množine. Avšak nemusíme zvoliť nič a teda aplikujeme štatistické metódy na celkovom množstve dát. Na obrázku 8.11 vidíme, že si môžme nastaviť časové rozmedzie ako aj interval rozdelenia daného časového okna. Okrem toho obsahujú všetky 3 dropdown-y

údaje z číselníkov, kde si môžeme zvoliť aj viac ako jednu možnosť.



Obr. 7.11: Úprava vstupov do analýzy

Ak spracovanie dát bolo úspešné, systém rovnako ako aj po prvotnom importe, zobrazí tento krát spracované dáta, na ktorých budeme vykonávať analýzu. Túto situáciu vidíme na obrázku 7.12. Okrem tabuľky spracovaných dát sú zobrazené po scroll-ovaní nižšie aj samotné výstupné sústavy grafov s popismi.

Obr. 7.12: Zobrazenie spracovaných dát

Výstupné sústavy grafov sú rozdelené na 2 časti. Prvé 3 grafy sú hlavné a odkazujú sa na dáta z tabuľky. Ostatné dáta sú pomocné histogramy, každý pre jeden typ filtra.

Obr. 7.13: Zobrazenie výstupných zostáv

Kapitola 8

Implementácia

V rámci tejto kapitoli budem rozoberať prostredie v ktorom som daný projekt implementoval, technológie, ktoré som pri práci použil, ďalej spôsob mapovania dát na relačnú databázu a vysvetlenie podstatných metód. V závere bude zhodnotenie otestovania bakalárskeho projektu ako aj popísané problémy s ktorými som narazil pri jeho implementácii.

8.1 Prostredie a technológie

Projekt som implementoval v prostredí *VS Code*. Webová aplikácia beží na localhoste a je zobrazovaná na webovom prehliadači Google Chrome. Z technológií som na prezenčnú vrstvu použil *HTML5, CSS, JavaScript a JQuery*. Databázovú vrstvu som zabezpečil pomocou *Postgresql* databázy, s ktorou som manipuloval prostredníctvom *pgAdmina verzie 4* a knižnice *psycopg2* pre *Python*. Aplikačnú logiku v zmysle komunikácie medzi vrstvami som implementoval v *Pythone*. Pre vytvorenie server client webovej aplikácie som použil *django framework*. Implementáciu štatistických metód a algoritmov analýz som robil pomocou knižníc numpy

a *matplotlib*. Na manipuláciu a parsovanie xml súborov som použil pymarc knižnicu a na parsovanie csv súborov csv knižnicu. Vypisovanie systémových logov v real-time asynchrónne som zabezpečil prostredníctvom framework-u *Ajax*.

8.2 Mapovanie dát

S dátami som pracoval prostredníctvom dátových štruktúr pythonu, zväčša to sú slovníky, listy, polia. Okrem toho boli hlavné dáta pre jednotlivé analýzy uložené v relačnej databáze. Na to som používal funkcionalitu *django framework-u* a síce ORM spôsob mapovania dát. Selekty z databáz boli aj vďaka tomu rýchle a jednoduché. O samotné mapovanie sa staral *models.py* súbor, kde boli zadefinované modely pre každú analýzu ako aj pomocné číselníky. Názvy hlavných modelov sú: *Analyza1Model, Analyza2Model, Analyza3Model*. Číselníky sú nasledovné: *TypOperacie, PscObvodu, TypKonspektu, VekovaSkupina*. Hlavné modely analýz sa skladajú z nasledovných atribútov:

1. Analýza Model

   - id = primárny kľúč
   - pouzivatelID = hash kód používateľa knižnice
   - vek = vek používateľa v čase vytvorenia transakcie
   - pohlavie = pohlavie používateľa
   - psc_id = PSČ, zároven aj kľúč do číselníka PscObvodu
   - casVytvoreniaTransakcie = date formát

2. Analýza Model

   - id = primárny kľúč

- transakciaId = identifikátor záznamu transakcie
- typOperacieň_id = typ transakcie, zároveň aj kľúč do číselníka TypOperacie
- casVytvoreniaTransakcie = date formát
- dlzkaVypozicky = dĺžka výpožičky v dňoch
- autor = meno autora vypožičaného titulu
- vydavateľstvo = názov vydavateľstva
- konspekt_id = číselný identifikátor a kľúč do číselníka Konspekt

3. Analýza Model

- id = primárny kľúč
- pouzivatelID = hash kód používateľa knižnice
- konspekt_id = číselný identifikátor a kľúč do číselníka Konspekt
- transakciaId = identifikátor záznamu transakcie

8.3 Implementácia aplikačnej logiky

Celá aplikačná logika bola naimplementovaná v jazyku *Python* s pomocou frameworku *django*. Django vytvorilo štruktúru projektu nasledovne: Celkové nastavenie server-client aplikácie je v *settings.py* súbore, kde som nastavil aj pripojenie na relačnú databázu. Routovanie, čiže smerovanie žiadostí na server a odpovedí riadi *urls.py*. Routovanie request-u obsahuje url cestu, odkaz na view (metódu, ktorá zabezpečuje renderovanie webového template-u) a názov daného routovania. V minulej sekcii som už spomenul *models.py*, ktorý zabezpečuje ORM modelovanie a komunikáciu s databázov. Z tých podstatných častí okrem html webových template-ov je súbor *views.py*. V danom súbore má každé routovanie svoju metódu,

ktorá zadáva obsah a renderuje template. Metóda index(request) je základnou view metódov s prázdnou url. Okrem metód slúžiacich na routovanie sa v tvomto súbore tvorí aj obsah a teda celková logika aplikácie. V nasledujúcich podsekciách vysvetlím funkcionalitu jednotlivých metód.

8.3.1 Import a validácia dát

- *def analyza(request)*: - prvá základná metóda, ktorá zabezpečuje importovanie dát z externých súborov. Metóda najprv kontroluje POST request, následne kontroluje prítomnosť súborov, validuje ich koncovku, volá metódy *csvSubor()* a *xmlSubor()*, ktoré zabezpečujú parsovanie dát z externých súborov, ich validáciu a následne naplnenie tabuľky údajmi pre konkrétnu analýzu. Okrem toho metóda zabezpečuje aj inicializáciu číselníkov. Táto metóda zároveň definuje obsah hlavného template-u na url *'/analyza'*, do ktorého metóda v *response* odošle obsah pre naplnenie tabuľky pre zvolenú analýzu.

- *def validujDataAnal1()*: - metóda *csvFile()* parsuje csv súbor po riadkoch. Každý riadok následne posiela validovať do tejto metódy pre Analýzu 1. Metóda je naimplementovaná na konkrétny typ csv súboru, čiže pri validácií automaticky kontroluje iba špecifické hodnoty v riadku na konkrétnej pozícii. Metóda konkrétny údaj kontroluje, či hodnota spĺňa daný formát, napríklad čas vytvorenia transakcie má istý formát a dĺžku. Metóda vracia slovník naplnení údajmi z daného riadka.

8.3.2 Spracovanie a agregovanie dát

- *def spracujVstupy(request, id)*: - po zadaní vstupov do analýzy užívateľom metóda príjme POST request ako svoj parameter. Jej úlohou je kontrolovať,

či sa pre konkrétny typ dát (napr dáta o vekovej skupine), nachádza vstup v requeste. Následne prostredníctvom ORM metóda vyselektuje všetky vyhovujúce hodnoty z tabuliek číselníkov podľa vstupných parametrov. Metóda uloží hodnoty do slovníka a vráti ich.

- *def spracovanieDat(vstupy, id)*: - po spracovaní vstupov, si dané vstupy metóda príjme ako parameter. Cieľom metódy je agregovať dáta na všetkých vstupoch naraz, čiže agregovať filtrovacie query, ktoré následne vyselektujú potrebné dáta z hlavnej tabuľky analýzy. Okrem toho metóda agreguje dáta aj jednotlivo pre konkrétne typy vstupov. Príkladom je agregovanie dát podľa vekovej skupiny, čiže selekt, ktorí rozdelí používateľov do vekových skupín a napočíta koľko sa v daných skupinách nachádza ľudí. Z toho vznikne histogram. Zaujímavou funkcionalitou je časové rozdelenie. Defaultný spôsob je porovnanie časového údaju prvej transakcie v záznamoch a posledného. Ak taktiež nie je zvolený interval, defaultný spôsob rozdelenia časového úseku je rozdelenie do dní. Následne metóda robí selekt na každý interval, čím vzniknú dvojice [dátum : počet transakcií používateľov], prípadne počet všetkých transakcií (aj opakujúcich sa vzhľadom na používateľa).

8.3.3 Použitie štatistických metód

- *def analyzaDat(vystupy)*: - Ako som už spomínal, tak produktom analýzy sú vždy 3 pohľady na hlavné agregované dáta pomocou metód regresie. Lineárna regresia, Polynomialna regresia a ku ním ešte histogram na finálnych dátach. Okrem nich metóda produkuje histogramy na vedľajších dátach. Základnou funkcionalitou použitia lineárnej a polynomiálnej regresie som sa inšpiroval na *w3schools* [27], prípadne na oficiálnych stránkach knižníc *matplotlib a numpy*.

8.4 Testy

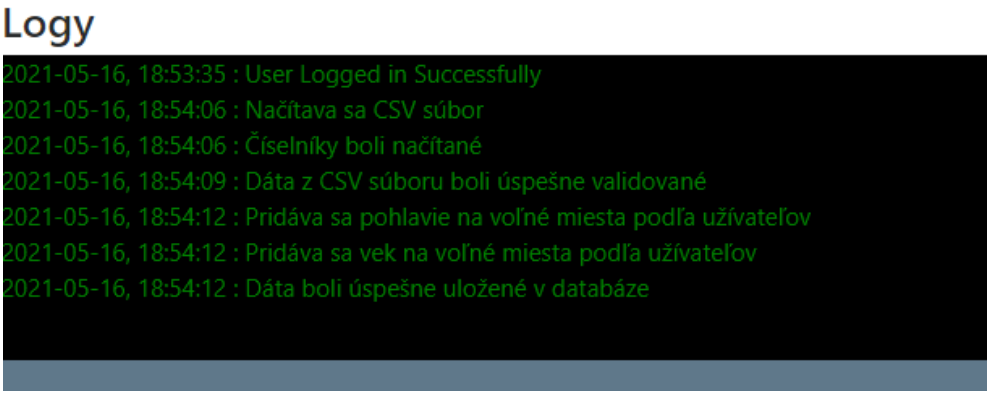
Na testovanie funkcionality systému som použil dataset z KIS České Budejovice [13], z ktorého som spravil tri datasety rôzneho rozsahu na porovnanie časových údajov zo spracovania a validovania dát. Prvý dataset obsahoval 5000 záznamov, ktoré pokrývali obdobie necelých 11 dní prevádzky systému v minulom roku v januári 2020. Druhý dataset obsahuje dvojnásobok záznamov, čiže 10000. Obdobie, ktoré pokrýva pokrýva celý január 2020. Tretím a najväčším datasetom je dataset záznamov za celý rok 2020. Obsahuje 112 000 tisíc záznamov. Porovnania datasetov na základe výkonu a správnosti riešenia som vykonal pomocou analýzy vývoja aktivity používateľov knižnice, v systéme ako Analýza 1.

8.4.1 Performance porovnanie

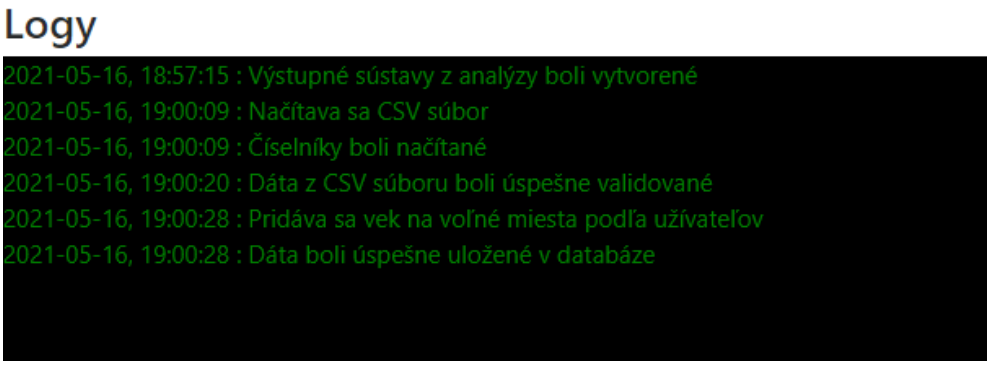
Na toto porovnanie som využil funkcionalitu logovania podstatných krokov v systéme. Asynchrónne real-time správy zabezpečil *Ajax*. Prvým datasetom bol dataset najmenšej veľkosti. Rozdieľ medzi časom importu csv súboru do systému a úspešným validovaním bol 3 sekundy. Rovnaký interval bol aj medzi validáciou a uložením záznamov do hlavnej tabuľky analýzy. Celkovo tak sledovaný časový úsek, pri veľkosti 5000 záznamov trval 6 sekúnd. Záznam z logov prvého datasetu je možné vidieť na obrázku 8.1.

Rovnaké testovanie časového úseku od importovania dát až po zápis do databázy som robil aj na veľkosťou stredne veľkom datasete. Dvojnásobný počet záznamov oproti prvému datasetu spôsobilo v prvom časovom úseku medzi importom a validáciou dát časový rozdiel 11 sekúnd. Druhý interval trval 8 sekúnd. Celkovo v porovnaní s prvým datasetom sa časy zhoršili v priemere trojnásobne. Záznam z logov druhého datasetu je možné vidieť na obrázku 8.2.

Počtom záznamov najväčším datasetom som očakával výraznejšie zhoršenie oproti



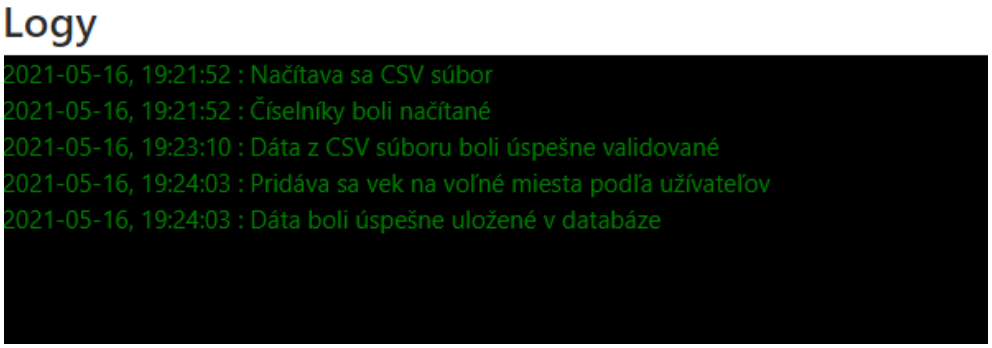
Obr. 8.1: Systémové logy spracovania najmenšieho datasetu



Obr. 8.2: Systémové logy spracovania prostredného datasetu

predošlým dvom výrazne menším datasetom. Najväčší dataset je oproti strednému datasetu vo veľkosti dát 11-krát väčší. Tento rozdieľ bol poznateľný v oboch časových úsekoch, ktorých trvanie bolo približne rovnaké. Prvý úsek medzi importom a validáciou trval 78 sekúnd. Druhý úsek trval o niečo menej a to 53 sekúnd. Pre porovnanie s druhým datasetom, boli časové údaje 7-krát väčšie. Záznam z logov tretieho datasetu je možné vidieť na obrázku 8.3.

Samozrejme po uložení dát v databáze nasledovalo ešte vypisovanie dát v tabuľke, ktoré sa stránkovalo po 5-tich záznamov. Časový úsek načítania dát z databázy trval v priemere do 10 sekúnd.

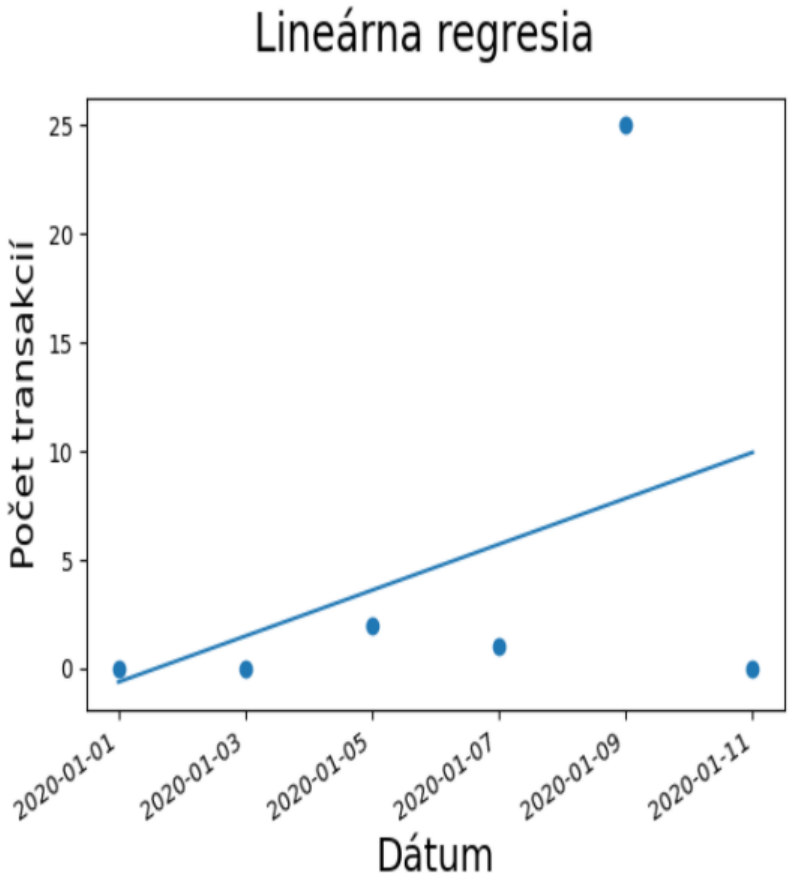


Obr. 8.3: Systémové logy spracovania najväčšieho datasetu

8.4.2 Porovnanie štatistických metód

Na všetkých datasetoch z porovnania výkonu som testoval aj funkcionalitu prvej analýzy. Cieľom analýzy bolo zistiť vývoj aktivít alebo systémových interakcií používateľov v zmysle vytvárania transakcií za konkrétne časové obdobie. Dáta o čitateľoch som agregoval rovnakým spôsobom na všetkých troch datasetoch. Testovacia vzorka tvorila záznamy všetkých žien, ktoré majú bydlisko v Českých Budejoviciach a ich vek spadá do vekovej skupiny od 31 rokov až po 50 rokov. Časový interval na ktorom som analyzoval testovaciu vzorku som rozdelil po dňoch. Výstupom pri všetkých analýzach boli okrem vedľajších histogramov aj 3 hlavné grafy. Konkrétne zobrazenie histogramu na testovacej vzorke a grafy lineárnej a polynomiálnej regresie.

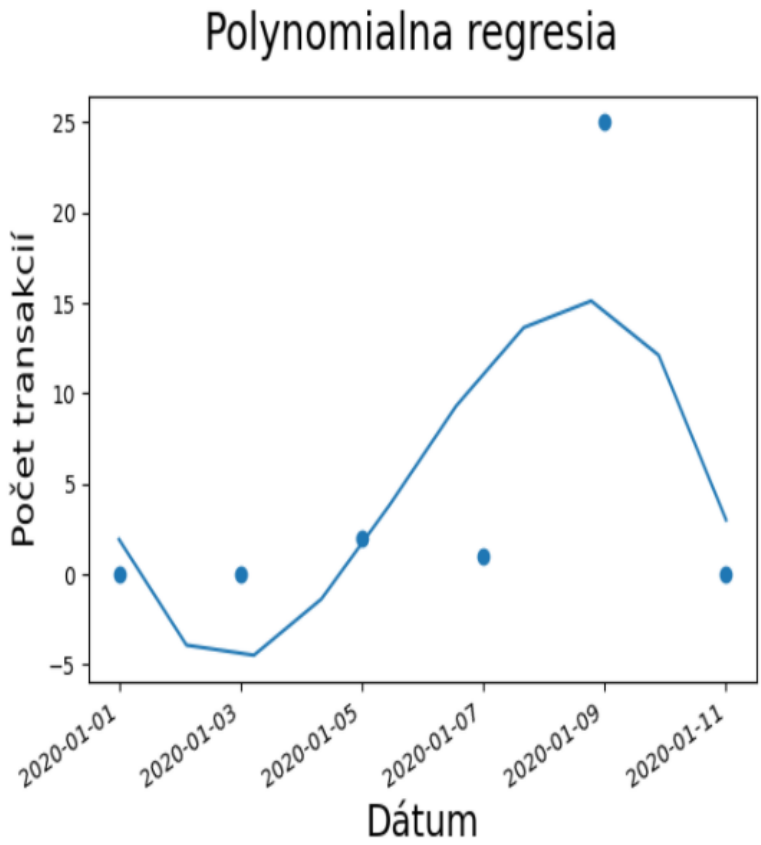
Keďže najmenší dataset pokrýva obdobie 11 dní, do mojej vzorky sa zmestila jedna žena. Jej aktivita počas prvých januárových dní bola až na jeden deň veľmi nízka. Na takejto testovacej vzorke vieme porovnať presnosť lineárnej a polynomiálnej regresie. Keďže dát bolo málo, lineárna regresia lepšie aproximovala dáta a teda kryvka sa vzhľadom na jednu vysokú hodnotu nevychýlila tak radikálne. Na obrázku 9.4 je možné vidieť graf lineárnej regresie a na obrázku 8.5 graf polynomiálnej regresie.



Obr. 8.4: Graf lineárnej regresie na vzorke z prvého datasetu

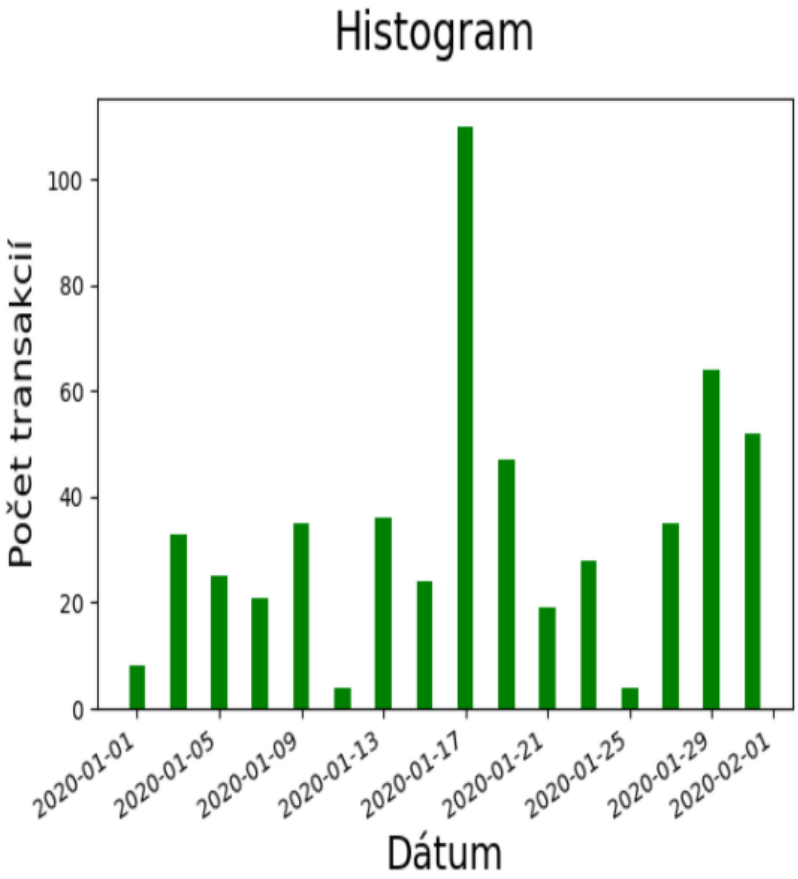
Prostredný dataset pokrýva obdobie jedného mesiaca a do mojej vzorky sa zmestili 4 ženy. Ich aktivita počas jedného mesiaca bola viac menej v stabilnom raste s pribúdajúcimi dňami, preto rozdieľ medzi lineárnov a polynomiálnom regresiou nebol tak značný. Obe metódy celkom presne aproximovali dáta. Na obrázku 8.6 je možné vidieť dáta reprezentované v histograme.

Tretí a najväčší dataset, ktorý pokrýva obdobie celého roka 2020, spôsobil veľmi reprezentatívnu vzorku aktivít 6-tich žien. Ich denná spoločná aktivita sa pohybovala vo veľkej časti aj v rádovo desiatkach transakcií denne. Na tejto vzorke je veľmi zaujímavý vývoj grafu polynomiálnej regresie, ktorá na rozdieľ od lineárnej

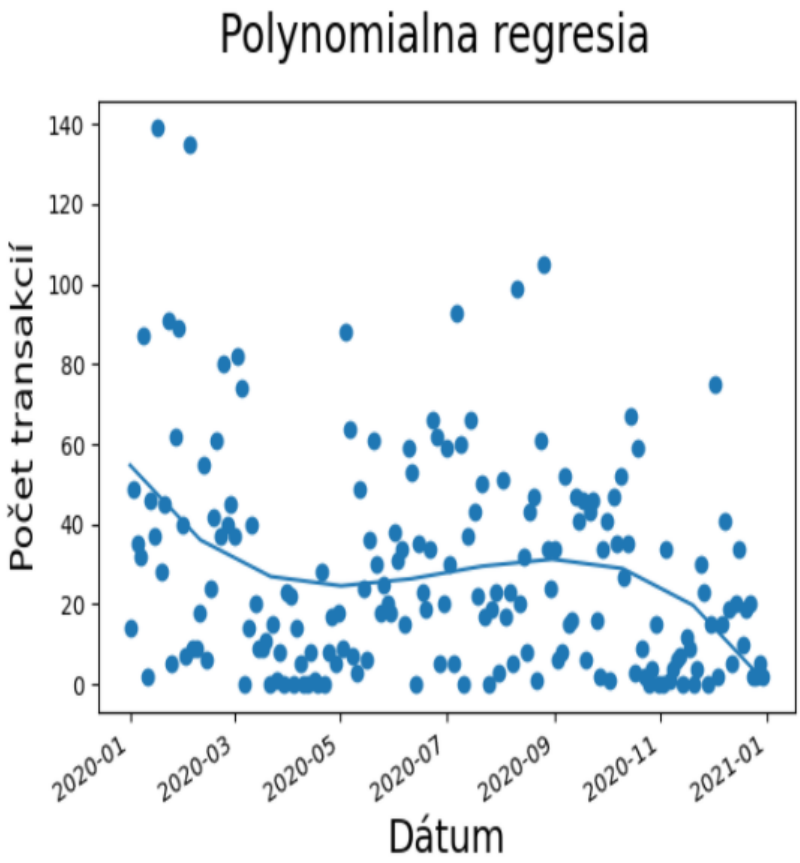


Obr. 8.5: Graf Polynomiálnej regresie na vzorke z prvého datasetu

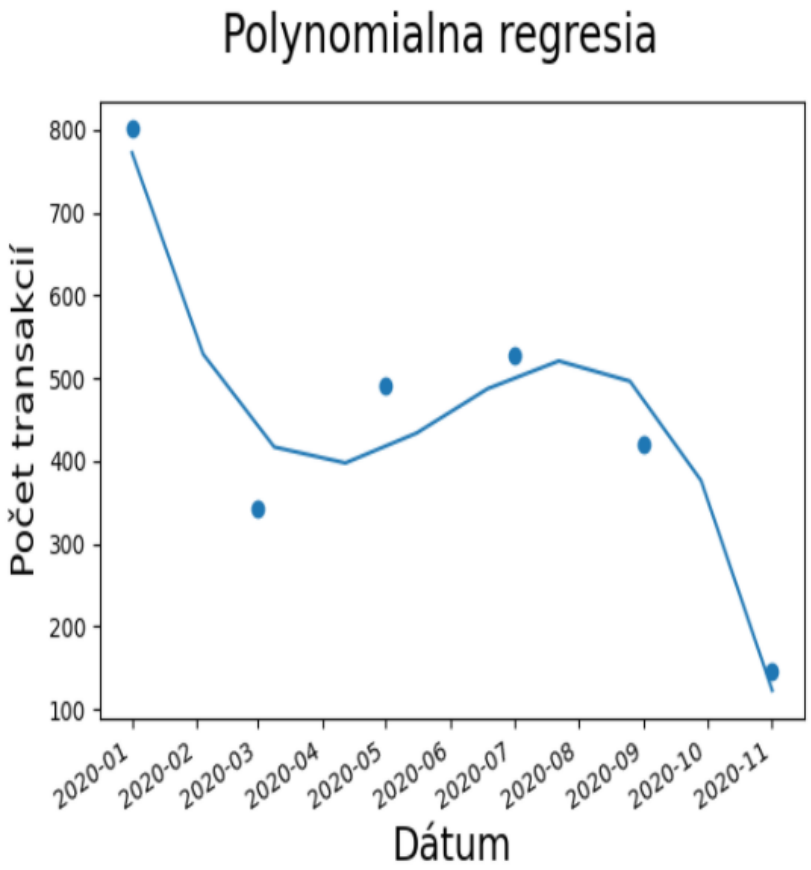
lepšie aproximovala dáta a ukázala aj zaujímavé trendy počas určitých období roka 2020. Zaujímavosťou je poznatok ako pandémia koronavírusu ovplyvnila v česku na jar a na jeseň návštevnosť knižníc. Na tomto datasete som následne aplikoval rovnakú testovaciu vzorku vstupov, ale čas jedného roka som rozdelil do mesiacov. Polynomiálna regresia na mesačných vzorkách v úvode bola mierne vychýlená, ale podstatnú časť roka aproximovala správne. Obidva grafy polynomiálnych regresií sú znázornené na obrázkoch 9.7 a 9.8.



Obr. 8.6: Histogram na vzorke z druhého datasetu



Obr. 8.7: Polynomiálna regresia na vzorke z tretieho datasetu



Obr. 8.8: Polynomiálna regresia na vzorke z tretieho datasetu rozdeléneho do mesiacov

8.5 Problémy

Počas implementácie som narazil na viacero menších problémov s nastavovaním prostredia alebo výberom správnych knižníc na aplikovanie štatistických metód. Hlavným problémom bola kvalita vstupných dát, ktorá následne spôsobila problémy v implementácií ich parsovania z csv súboru. Csv súbor, kde boli všetky záznamy transakcií som čítal po riadkoch a rozdeľoval údaje v riadkoch podľa čiarok. Prvým problémom bolo prítomnosť viacerých čiarok ako bol počet údajov v riadku, čo spôsobovalo nesprávne parsovanie a následne chyby pri validovaní dát. Druhým problémom bola kvalita dát v zmysle nekonzistentnosti ich formátu. Napríklad údaj o cene bol reprezentovaný viacerými spôsobmi, kedy boli použité aj čiarky aj bodky na oddelenie desatinných miest alebo údaje o poštových smerovacích číslach boli reprezentované buď priamo číselne alebo aj názvom daného obvodu. Číselné hodnoty boli s medzerou aj bez medzery. Pri niektorých záznamoch boli určité hodnoty prázdne. Pre konzistetnosť a použiteľnosť som ich nahrádzal náhodnými hodnotami. Takýto spôsob náhodného dopĺňania dát nie je vhodný pre produkčné použitie systému a prácu s reálnymi dátami a to pre skreslené výsledky analýz, avšak pre overenie správnej funkcionality kódu to bolo postačujúce.

Kapitola 9

Záver

Cieľom bakalárskej práce bolo analyzovať typy dát v knižniciach alebo ich informačných systémoch a analyzovať dostupné štatistické metódy, ktorými by bolo vhodné analyzovať dané dáta. Navrhnutý informačný systém reprezentuje nadstavbu pre knižnično-informačné systémy a môže poskytovať zamestnancom knižnice variabilitu pri tvorbe výstupných zostáv, štatistík a rôznych pohľadov na dáta z ich informačných systémov so zameraním na odhaľovanie nových poznatkov. Návrh systému hlavne spočíval v správnej validácií testovacích dát, ktorá bola základom pre správne namapovanie na konkrétne typy analýz. Variabilita bola navrhnutá tak aby zamestnanec knižnice mal na výber akým spôsobom chce agregovať testovacie dáta a tým viac konkretizovať analýzu na užšej špecifickej vzorke. Použitím korelačných štatistických metód systém vytvára výstupne zostavy v podobe grafov a tabuliek, ktorými odhaluje a predikuje vývoj interakcií medzi používateľmi knižnice a knižnicou na konkrétnom časovom úseku. Systém bol navrhnutý tak aby bol ľahko rozšíriteľný o ďalšie typy analýz na dátach z knižníc. Implementácia systému odhalila niekoľko problémov pri parsovaní dát sposobených kvalitov vstupných testovacích dát. Z navrhovaných analýz bola im-

plementovaná analýza vývoja aktivity používateľov prostredníctvom transakcií v knižniciach. Testovaním danej analýzy na rozsiahlych testovacích vzorkách s možnosťou veľkej variability pri ich agregovaní sa tak overila naimplementovaná funkcionalita. Výstupné zostavy nebolo možné porovnať s reálnymi dátami, preto sa v rámci testovania porovnávali výsledky použitia lineárnej a polynomiálnej regresie. Zlepšením systému do budúcnosti by bolo implementovanie ďalších typov analýz, rozšírením variability tvorby výstupných zostáv napríklad priamym upravovaním parametrov štatistických metód ako napríklad stupňa polynomiálnej funkcie. Zlepšením by bola aj možnosť importovať dáta prostredníctom OAI-PMH.

Literatúra

[1] URL: http://www.unesco.sk/nehmotne-kulturne-dedicstvo-SR.

[2] URL: https://www.zakonypreludi.sk/zz/2015-126.

[3] URL: https://www.svop.eu/index.php/produkty/kis-dawinci/popis-modulov/protokol-z39-50.

[4] URL: https://www.svop.eu/index.php/produkty/kis-dawinci.

[5] URL: https://www.svop.eu/index.php/produkty/kis-dawinci/popis-modulov/modul-vypozicky/110-vypozicny-modul.

[6] URL: https://www.librarianshipstudies.com/2018/12/anglo-american-cataloguing-rules-aacr.html.

[7] URL: https://www.ulib.sk/files/SKP/MARC21bib.pdf.

[8] URL: https://aleph.nkp.cz/F/?func=direct&doc_number=000001371&local_base=KTD.

[9] URL: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2002/428/20050101.html.

[10] URL: https://www.epi.sk/zz/2016-201.

[11] URL: http://www.kniznicatopolcany.sk/images/documents/kniznicny_a_vypozicny_poriadok_2020.pdf.

[12] URL: https://www.snk.sk/images/sluzby/sidelna_budova/KP_2007.pdf.

[13] URL: https://arl2.library.sk/transfer/mj/outstud.zip.

[14] URL: http://www.kniznicatopolcany.sk/images/documents/Sprava_o_cinnosti_TK_za_rok_2019.pdf.

[15] URL: https://www.culture.gov.sk/wp-content/uploads/2020/12/KULT-10-01_o_kniznici_2020.pdf.

[16] URL: https://www.altmetric.com/blog/assessing-digital-library-metrics/.

[17] URL: https://www.loc.gov/marc/classification/cdintro.html.

[18] URL: https://www.loc.gov/standards/mods/design-principles-mods-mads.html.

[19] URL: https://www.loc.gov/standards/mods/mods-outline-3-7.html.

[20] URL: https://www.loc.gov/standards/marcxml/.

[21] URL: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

[22] URL: http://www.dlib.org/dlib/july03/young/07young.html.

[23] URL: https://www.hindawi.com/oai-pmh/.

[24] URL: https://www.sciencedirect.com/science/article/pii/S2212827119307401.

[25] URL: https://www.ibm.com/support/knowledgecenter/SSFMBX/com.ibm.swg.im.dashdb.analytics.doc/doc/r_statistics.html.

[26] URL: https://content.sciendo.com/view/journals/dim/2/2/article-p103.xml?language=en.

[27] URL: https://www.w3schools.com/python/python_ml_polynomial_regression.asp.

[28] Admin. *Knižnično-informačný systém Advanced Rapid Library (ARL)*. URL: http://www.uk.sav.sk/uk_pre_SAV/pre-knihovnikov/2014/10/09/kniznicno-informacny-system-advanced-rapid-library-arl/.

[29] Aman Bhardwaj. *Online Library Management System*. URL: https://www.academia.edu/8756988/Online%7B%5C_%7DLibrary%7B%5C_%7DManagement%7B%5C_%7DSystem (cit. 22. 10. 2019).

[30] Rafael Ball. "Big Data and Their Impact on Libraries". In: *American Journal of Information Science and Technology* 3.1 (2019), s. 1. DOI: 10.11648/j.ajist.20190301.11.

[31] *Deklarácia*. URL: http://www.culture.gov.sk/posobnost-ministerstva/kulturne-dedicstvo-/ochrana-pamiatok/dokumenty/deklaracia-1bb.html.

[32] Dr. Azeez Ahmed Dr. Kousar Jaha Begum. "The Importance of Statistical Tools in Research Work". In: (), s. 1–9. URL: https://www.arcjournals.org/pdfs/ijsimr/v3-i12/10.pdf.

[33] Anand Gupta et al. "A Big Data Analysis Framework Using Apache Spark and Deep Learning". In: *IEEE International Conference on Data Mining Workshops, ICDMW*. Zv. 2017-Novem. IEEE Computer Society, dec. 2017, s. 9–16. ISBN: 9781538614808. DOI: 10.1109/ICDMW.2017.9. arXiv: 1711.09279.

[34] *Hlavná stránka*. URL: https://sek.euba.sk/o-kniznici/kniznicny-fond.

[35] Jennifer Knievel. *Use of circulation statistics and interlibrary loan data in collection management*. Tech. spr. 2006, s. 7–16. URL: https://core.ac.uk/download/pdf/54848054.pdf.

[36] Ana Kovacevic, Vladan Devedzic a Viktor Pocajt. "Using data mining to improve digital library services". In: *Electronic Library* 28.6 (2010), s. 829–843. DOI: 10.1108/02640471011093525.

[37] Jian Wei Li a Ping Hua Chen. "The application of cluster analysis in library system". In: *2008 IEEE International Symposium on Knowledge Acquisition*

and Modeling Workshop Proceedings, KAM 2008. 2008, s. 907–910. ISBN: 9781424435296. DOI: 10.1109/KAMW.2008.4810639.

[38] *Library classification*. Dec. 2019. URL: https://en.m.wikipedia.org/wiki/Library_classification.

[39] *Library management*. Nov. 2019. URL: https://en.wikipedia.org/wiki/Library_management.

[40] Ayushi Malviya, Amit Udhani a Suryakant Soni. "R-Tool: Data analytic framework for big data". In: *2016 Symposium on Colossal Data Analysis and Networking, CDAN 2016*. Institute of Electrical a Electronics Engineers Inc., sept. 2016. ISBN: 9781509006694. DOI: 10.1109/CDAN.2016.7570960.

[41] Yongming Wang Jia Mi. "Applying Statistical Methods to Library Data Analysis". In: (2019), s. 1–7. URL: https://www.tandfonline.com/doi/pdf/10.1080/0361526X.2019.1590774?needAccess=true.

[42] International Conference On Computation Of Power, Energy Information a And Communication. *STUDY OF CLUSTERING ALGORITHMS FOR LIBRARY MANAGEMENT SYSTEM*. 2018. ISBN: 9781538624470.

[43] Slovensk. *Zbierkov fond SNM*. URL: https://www.snm.sk/?zaujimavosti-zbierkovych-fondov.

[44] Michal Švec. *Knižničný fond*. URL: https://www.uk.ukf.sk/sk/kniznicny-fond.

[45] Chunning Wang et al. "Exposing library data with big data technology: A review". In: *2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 - Proceedings*. Institute of Electrical a Electronics Engineers Inc., aug. 2016. ISBN: 9781509008063. DOI: 10.1109/ICIS.2016.7550937.