Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.42
DOI: https://doi.org/10.24867/10BE35Adamov

SISTEM ZA SPAJANJE BAZA PODATAKA BIBLIOTEČKOG INFORMACIONOG SISTEMA

SYSTEM FOR MERGING LIBRARY INFORMATION SYSTEM DATABASES

Ivan Adamov, *Fakultet tehničkih nauka, Novi Sad*

Oblast – SOFTVERSKO INŽENJERSTVO I INFORMACIONE TEHNOLOGIJE

Kratak sadržaj – *Predmet ovog rada jeste opis sistema za spajanje baza podataka bibliotečkog informacionog sistema BISIS. U prvom delu rada opisan je bibliotečki informacioni sistem BISIS i UNIMARC format. U nastavku je opisana specifikacija sistema, implementacija i rad sistema. Opisana su dva režima rada sistema: pun i inkrementalni. Na kraju rada su opisani dobijeni rezulatati i performanse izvršavanja sistema.*

Ključne reči: *MongoDB, BISIS, spajanje*

Abstract – *The subject of this paper is a description of the system for merging databases of the library information system BISIS. The first part of the paper describes the library information system BISIS and UNIMARC format. The system specification, implementation and operation of the system are described below. Two operating modes of the system are described: full and incremental. At the end of the paper, the obtained results and system execution performance are described.*

Keywords: *MongoDB, BISIS, merge*

1. UVOD

Bibliotečki informacioni sistem BISIS [1] predstavlja sistem za bibliotečko poslovanje. Trenutno obuhvata više od 50 biblioteka širom Srbije, kao i jednu u Austriji, u Lincu. Bibioteke su različitog tipa: javne, visokoškolske, školske i specijalizovane. Podaci ovih biblioteka smeštaju se u zasebne baze podataka.

U okviru ovog sistema postoji sistem za pretragu bibliotečke građe. Kada korisnik unese neki upit, sistem pretražuje zasebne baze podataka i korisniku prikazuje rezultat pretrage. Kako bi se poboljšalo funkcionisanje sistema i izbeglo pretraživanje zasebnih baza podataka, trebalo je sve te podatke spojiti u jednu bazu padataka.
Tema ovog rada jeste implementacija algoritma za spajanje više baza podataka u jednu bazu podataka. Kao testne baze korišćene su baze sledećih biblioteka:
Biblioteka grada Beograda, Gradska biblioteka Novog Sada, Biblioteka šabačka, Biblioteka Milutin Bojić.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Branko Milosavljević, red. prof.

1.1. Bibliotečki informacioni sistem

Godine 2017. započet je razvoj pete verzije bibliotečkog informacionog sistema BISIS u saradnji sa Bibliotekom grada Beograda, Gradskom bibliotekom u Novom Sadu i Bibliotekom šabačkom. Uvedeni su savremeni tehnološki standardi u bibliotekarstvo u Srbiji, pojednostavljeno je održavanje sistema i omogućena je masovnija primena.

1.2. Najvažnije karakteristike

Najvažnije karakteristike bibliotečkog informacionog sistema BISIS verzije 5.0 [2] su:

- Katalogizacija - potpuno u skladu sa međunarodnim standardom UNIMARC.
- Cirkulacija - cirkulacija (pozajmica) fonda prema praksi srpskih najrazvijenijih biblioteka.
- Izveštavanje - detaljno izveštavanje o građi, pozajmici i drugim aktivnostima.
- Nabavka - automatizovan proces nabavke knjiga povezan sa katalogizacijom.
- OPAC (*online public access catalog*) - pretraživanje fonda za korisnike.
- Objedinjeni katalog - objedinjeni katalog svih biblioteka u sistemu BISIS.
- Digitalni sadržaji - mogućnost skladištenja i pregleda digitalnih sadržaja.
- Uzajamna katalogizacija - preuzimanje zapisa iz drugih biblioteka u sistemu BISIS.
- Otvoreni kôd - kôd BISIS sistema je dostupan svima

1.3. UNIMARC format

UNIMARC (*Universal MARC (Machine Readable Cataloging) format*) [3] je format za obeležavanje i razmenu bibliografskih podataka. Jedan zapis se sastoji od više polja. Svako polje sadrži UNIMARC kôd koji predstavlja identifikator polja i sastoji se od tri cifre. Polja sadrže i podpolja koja se sastoje od identifikatora podpolja i sadržaja podpolja. Identifikator podpolja je predstavljen jednim slovom ili cifrom. Primer kodova UNIMARC formata korišćenih u BISIS sistemu prikazan je u tabeli 1.

Kod polja	Kod podpolja	Značenje
010	a	ISBN
011	a	ISSN
101	a	Jezik knjige
102	a	Zemlja izdavanja
200	a	Naslov knjige
200	e	Podnaslov
700	a	Prezime prvog autora
700	b	Ime prvog autora
210	a	Mesto izdavanja
210	c	Izdavač
210	d	Godina izdavanja
215	a	Broj stranica

Tabela 1. *primer UNIMARC kodova*

2. KORIŠĆENE TEHNOLOGIJE

Aplikacija je napisana u Java programskom jeziku. Podaci bibliotečkog informacionog sistema BISIS čuvaju se u MongoDB [4] bazi podataka. Pored MongoDB baze korišćena je i Redis [5] baza podataka koja služi za povećanje performansi izvršavanja algoritma.

2.1. MongoDB

MongoDB je NoSQL [6] (nerelaciona) baza podataka koja podatke čuva u vidu dokumenata u BSON (*Binary JSON*) [7] formatu koji u velikoj meri liči na JSON (*JavaScript Object Notation*) [8] format. U MongoDB se koriste pojmovi: baza podataka, kolekcija i dokument. Baza podataka može imati nula ili više kolekcija. Jedna kolekcija može imati nula ili više dokumenata. Kolekcija ne određuje šemu baze podataka i ekvivalentna je tabeli u relacionim bazama podataka. Dokument predstavlja jedan zapis u BSON formatu. Dokument u kolekciji ekvivalent je jednom redu u tabeli u relacionoj bazi podataka. Polje dokumenta je isto što i kolona kod relacione baze. Svaki dokument ima jedinstveni identifikator koji je tipa *ObjectId.*

2.1.1. JSON i BSON formati

JSON je tekstualni format i najčešće se upotrebljava za komunikaciju na vebu. Jednostavan je i lako se čita.

BSON predstavlja JSON u binarnom formatu. Kasnije je proširen novim tipovima podataka kako bi poboljšao određene nedostatke JSON formata koji u velikoj meri povećavaju mogućnosti pretrage nad MongoDB bazom.

Kroz dokumente u BSON formatu brže se prolazi nego kroz dokumete u JSON formatu. BSON ima brzu serijalizaciju i deserijalizaciju i zauzima više prostora od JSON-a jer čuva dodatne podatke kao što je dužina polja, što omogućava brz prolazak kroz dokument.

2.2. Redis

Redis je veoma brza, NoSQL, *in-memory* baza podataka i najčešće se koristi za keširanje podataka. Podaci se čuvaju u obliku ključ-vrednost i moguće je podesiti vremenski interval čuvanja podataka (posle isteka vremenskog intervala podaci se brišu). Kao vrednost se mogu staviti razni tipovi podataka kao što su: string, lista, mapa, set, sortirani set, i drugi. Redis je projekat otvorenog koda i dostupan je svima.

3. SPECIFIKACIJA SISTEMA

Podaci o bibliotečkim građama pojedninačnih biblioteka BISIS sistema čuvaju se u zasebnim bazama podataka. BISIS sistem obuhvata i takozvani OPAC, sistem za pretraživanje bibliotečkih građa. Taj sistem omogućava pretraživanje bibliotečkih građa po: naslovu, autoru, ključnim rečima, izdavaču i godini izdavanja uz mogućnost dodavanja logičkih operatora: *i, ili, ne*. Kada korisnik sistema unese neki tekst i klikne *enter*, izvršavaju se sledeći koraci:

1. Sistem u pozadini pretražuje određeni Elasticsearch indeks [9].

2. Sistem dobija rezultat u kojem je navedeno u kojim sve bazama se nalazi zadati tekst.

3. Sistem šalje nove upite ka pojedinačnim bazama iz dobijenog rezultata iz koraka 2.

4. Sistem dobija podatke iz pojedinačnih baza i vraća rezultat krajnjem korisniku.

Cilj ovog rada je poboljšanje rada sistema za pretragu, tačnije poboljšanje koraka 3. Umesto da sistem šalje zahteve ka pojedinačnim bazama podataka, treba napraviti jednu bazu koja će čuvati sve podatke bibliotečkog informacionog sistema. Cilj zadatka je implementacija algoritma za spajanje baza podataka u dva režima:

1. Pun režim - spajanje baza podataka “od nule”. Predstavlja spajanje svih podataka pojedinačnih baza u jednu bazu unije.

2. Inkrementalni režim - spajanje novokreiranih podataka iz pojedinačnih baza sa unija bazom.

Kao testne baze koristiti baze sledećih biblioteka: Biblioteka grada Beograda (BGB), Gradska biblioteka Novog Sada (GBNS), Biblioteka šabačka (BS), Biblioteka Milutin Bojić (BMB). Zapise treba sačuvati u bazu unije (UNION). Za kriterijum poređenja koristiti:

1. ISBN

2. ISSN

3. Naslov, autor, izdavač, godina izdavanja

U rezultat ulaze podaci baza i to redom po prioritetu (ako se jedna bibliotečka građa nalazi u dve baze i zapisi u tim bazama imaju različite vrednosti za neko polje, uzeće se vrednost polja iz baze koja ima viši prioritet):

1. BGB

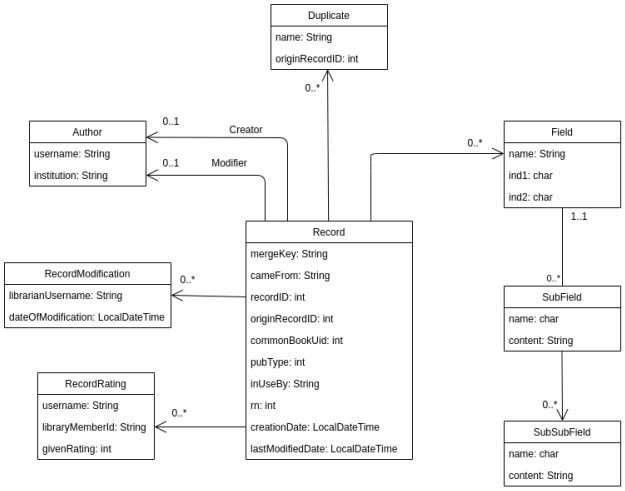
2. GBNS

3. BS

4. BMB

3.1. Dijagram klasa

Klasa *Record* opisuje bibliotečku građu kao i dodatne informacije vezane za pojedinačan zapis (slika 1).



Slika 1. *Dijagram klasa*

4. IMPLEMENTACIJA SISTEMA

Algoritam je implementiran u Java programskom jeziku. Algoritam radi u dva režima: pun i inkrementalni.

4.1. Pun režim

Pun režim predstavlja spajanje baza podataka biliotečkog informacionog sistema od "nule", tj. spajanje svih podataka u jednu bazu unije. Spajanje zapisa se vrši prvo po ISBN-u, zatim po ISSN-u i na kraju po naslovu bibliotečke građe.

Sistem prvo preuzima sve zapise iz pojedinačnih biblioteka koji imaju ISBN, zatim sve zapise koji imaju ISSN, a nemaju ISBN, i na kraju sve zapise koji imaju naslov, a nemaju ISBN i ISSN. Posle preuzimanja svake grupe zapisa radi se spajanje.

Prilikom spajanja baza kada neka bibliotečka građa već postoji u *union* bazi, treba da se uradi ažuriranje tog zapisa. Potrebno je imati Java objekat zapisa da bi se uradile određene operacije poređenja i spajanja dva objekta. Dakle, potrebno je uraditi novi upit ka bazi i dobaviti određeni zapis. Umesto da se upit šalje ka *union* (Mongo) bazi, u Redis bazu se smeštaju svi zapisi koji se upišu u *union* bazu. Ovim načinom se dobija veliko ubrzanje rada algoritma.

Bibliotečka građa se jedinstveno identifikuje ključem zapisa i služi za proveravanje da li bibliotečka građa već postoji u *union* bazi. Primer ključa zapisa za bibliotečku građu "Seobe" dat je na listingu 1.

```
8619018086#_#seobe#nolit#crnjanski milos#1990
```

Listing 1. *Primer ključa zapisa*

Ključ zapisa se sastoji od:

1. Transformisani ISBN, ISBN bez "-" znakova i ako je 13-ocifreni broj izbačena su prva tri broja (978).

2. Transformisani ISSN, ISSN bez "-" znakova.

3. Tranformisani naslov, naslov pretvoren u mala slova latinicom.

4. Tranformisano ime izdavača, ime izdavača pretvoreno u mala slova latinicom.

5. Tranformisano ime autora, ime autora pretvoreno u mala slova latinicom.

6. Godina izdavanja.

7. Separator "#" - služi da odvoji vrednosti.

8. Prazno polje "_" - ako bibliotečka građa ne sadrži neku od vrednosti (ISBN, ISSN, naslov, ...) stavlja se znak "_".

Prilikom pokretanja punog režima, *union* baza je prazna. Prvi korak predstavlja dodavanje BGB zapisa, jer BGB baza ima najveći prioritet. Dodavanje velike količine podataka u Mongo bazu najbrže se izvršava u *batch* režimu. *Batch* režim predstavlja dodavanje zapisa u grupama, npr. dodavanje 1000 zapisa odjednom. Postoje duplikati istih zapisa u pojedinačnim bazama bibliotečkog informacionog sistema. Ovaj sistem vodi računa da se isti zapisi iz istih baza ne upisuju više puta u *union* bazu. BGB zapis koji treba da se sačuva u *union* bazu proširuje se dodatnim podacima kao što su: id zapisa u originalnog bazi i ključ zapisa. Zapis se dodaje u *batch* kontejner i JSON zapis se smešta u Redis bazu. Kada se skupi određen broj zapisa, zapisi se upisuju u *union* bazu u *batch* režimu.

Nakon dodavanja BGB zapisa dobavljaju se zapisi iz ostalih baza (GBNS, BS, BMB) po definisanom prioritetu. Za svaki dobavljeni zapis se generiše ključ zapisa i na osnovu tog ključa se radi provera da li bibliotečka građa već postoji u *union* bazi. Ako zapis ne postoji, dodaće se isto kao i BGB zapis, a ako postoji uradiće se ažuriranje zapisa.

Zapis koji se već nalazi u *union* bazi se dobavlja iz Redis baze na osnovu ključa zapisa i radi se spajanje zapisa iz baze manjeg prioriteta sa zapisom u *union* bazi. Umesto da se radi klasično ažuriranje (*update* upit), prvo se izbrišu svi zapisi koji treba da se ažuriraju i zatim se dodaju ažurirani zapisi. Na ovaj način se postiže veliko ubrzanje rada algoritma.

4.2. Inkrementalni režim

Inkrementalni režim predstavlja spajanje novokreiranih zapisa sa *union* bazom. Zamišljeno je da se pokreće periodično (npr. jednom nedeljno). Kod inkrementalnog režima se umesto pretraživanja zapisa po ISBN-u, ISSN-u i naslovu pretražuju svi zapisi čije je vreme kreiranja veće od vrednosti poslednjeg pokretanja ovog režima rada. Prilikom dodavanja novog zapisa u *union* bazu u punom ili inkrementalnom režimu za svaki zapis se čuva i njegov ključ zapisa. U inkrementalnom režimu se na osnovu ključa zapisa proverava da li zapis postoji u *union* bazi. Ako ne postoji, treba da se doda, a ako postoji, treba da se ažurira. Sistem ima zaštitu od spajanja istog zapisa više puta sa *union* bazom. Kada se završi spajanje snima se poslednje vreme kada je izvršeno spajanje. Poslednje vreme spajanja čuva se u običnom tekstualnom fajlu.

Za razliku od punog režima u ovom režimu rada ne koristi se Redis baza podataka, a provera da li zapis već postoji radi se tako što se šalje upit ka Mongo bazi, dok se ažuriranje zapisa radi na klasičan način (*update* upit).

5. PERFORMANSE

Bitna karakteristika algoritma je i brzina izvršavanja. Prilikom prve implementacije sistemu je bilo potrebno od 10 do 15 dana da spoji četiri baze podataka, da bi se nakon toga u *union* bazi nalazilo preko 400.000 zapisa. Određenim poboljšanjima došlo se do konačne i trenutne verzije algoritma koja četiri baze podataka spoji za oko 220 sekundi.

Kod je bio podeljen na logičke celine i onda se vršilo merenje vremena izvršavanja i analiza logičkih celina. Prvobitno se dodavanje i ažuriranje zapisa radilo jedan po jedan. Uvođenjem *batch* režima postiglo se značajno ubrzanje. Proveravanje postojanja zapisa u *union* bazi kod punog režima se radilo tako što se slao kompleksni upit ka *union* (MongoDB) bazi. Uvođenjem Redis baze podataka i keširanjem dodatih zapisa brzina rada algoritma povećala se oko 22 puta. Ažuriranje zapisa slanjem *update* upita izvršavalo se značajnije sporije nego korišćenjem hibridnog pristupa. Svi zapisi koje je potrebno ažurirati čuvaju se u nekoj listi, zatim se prvo izbrišu iz *union* baze, pa se dodaju. Dakle umesto *update* upita koriste se *delete* + *insert*. Ovaj pristup je doneo veliko ubrzanje rada algoritma.

6. ZAKLJUČAK

Bibliotečki informacioni sistem BISIS predstavlja veoma značajno unapređenje bibliotečkog poslovanja. Sistem se unapređivao i unapređivaće se kako godine budu odmicale. Sistem za spajanje baza podataka predstavlja jedan korak u unapređenju rada sistema.

Početni korak prlikom razvoja sistema bio je analiza tehnologija koje se koriste u bibliotečkom informacionom sistemu kao i analiza karakteristika UNIMARC formata koji služi za reprezentaciju bibliografske građe. U radu su opisane tehnologije za razvoj sistema za spajanje baza podataka, kao i UNIMARC format sa datim primerima. Data su poređenja, prednosti i mane JSON i BSON formata podataka. Naredni korak bio je razvoj samog sistema. Sistem omogućava formiranje jedinstvene baze UNIMARC zapisa. Korišćenjem određenih metoda uočeni su zapisi koji predstavljaju identična izdanja i implementiran je algoritam za spajanje tih zapisa. Prilikom spajanja baza ispoštovana su pravila prioriteta baza. Poređenje zapisa je *case* i *script insensitive*, dakle podaci se svode na isti format - "ošišanu" latinicu malim slovima. ISBN i ISSN formati su svedeni na isti format - desetocifreni broj bez crtica.

Nakon završetka inicijalne verzije sistema, sistem u punom režimu je radio dobro, ali sporo. Velika ubrzanja postignuta su keširanjem podataka i uvođenjem Redis baze podataka, korišćenjem *batch* režima rada i hibridnog ažuriranja (*delete* + *insert* umesto *update* upita). Krajnja verzija sistema radi dobro i brzo. Sistem je otporan na loše podatke, dobro uočava duplikate i lako ga je proširiti na rad sa više baza podataka.

Dalji pravci razvoja sistema se ogledaju u proširivanju sistema na rad sa više baza podataka i optimizaciji zauzeća memorije koje zauzimaju keširani podaci.

7. LITERATURA

[1] https://bisis.rs/biblioteke.html

[2] https://bisis.rs

[3] https://www.ifla.org/publications/unimarc-formats-and-related-documentation

[4] https://en.wikipedia.org/wiki/MongoDB

[5] https://en.wikipedia.org/wiki/Redis

[6] https://en.wikipedia.org/wiki/NoSQL

[7] https://en.wikipedia.org/wiki/BSON

[8] https://www.json.org

[9]https://www.elastic.co/blog/what-is-an-elasticsearch-index

Kratka biografija:

Ivan Adamov rođen je 30.06.1996. u Zrenjaninu. Srednju Elektrotehničku i građevinsku školu "Nikola Tesla" završio je 2015. godine u Zrenjaninu. Iste godine upisuje Fakultet tehničkih nauka u Novom Sadu, smer Softversko inženjerstvo i informacione tehnologije. Studije završava 2019. godine. Iste godine upisuje master akademske studije na istom fakultetu, smer "Elektronsko poslovanje". Master rad odbranio je 2020. godine.