<KIV> KATEDRA INFORMATIKY A VÝPOČETNÍ TECHNIKY

Vojtěch Danišík

Diplomová práce

Inženýrská informatika
Softwarové inženýrství
2021/2022

Vedoucí práce:
Doc. Ing. Dalibor Fiala, PhD.

Tvorba rozsáhlých úložišť patentových dat

Abstrakt

Předmětem této diplomové práce je seznámit se s dostupnými zdroji dat o patentech a vytvořit rozsáhlá lokální úložiště patentových dat umožňující jejich efektivní prohledávání a vytěžování. V první části posteru byly popsány typy patentů a lokální úložiště. V druhé části posteru byl popsán výběr patentových dat, výběr databází pro lokální úložiště a jejich realizace.

Úvod

Hlavním cílem této práce je vytvoření lokálního úložiště, které bude umožňovat efektivní vytěžování patentových dat stažených z dostupných databází patentových institucí. Stažená data budou filtrována na základě specifikovaných atributů, které musí obsahovat.

Východiska, analytická část

Patent je grant od vlády, který dává vynálezci právo vyloučit ostatní z výroby, používání, prodeje dovozu nebo nabízení vynálezu pro prodej na dobu určitou. Patenty jsou klasifikovány do 3 hlavních typů:

- Užitný patent—patent na užitečný proces, výrobní předmět, stroj a jiné
- Návrhový patent—patent na originální, nové a ornamentální vzory pro vynálezy
- Patent rostlin—patent na novou nebo odlišnou odrůdu rostliny

Databáze je termín, který označuje organizovanou kolekci strukturovaných informací nebo dat, která jsou typicky ukládána elektronicky v počítačovém systému.

Z mnoha typů databází byly vybrány a analyzovány (vzhledem k typu dat) tyto typy databáze: relační, grafová, objektově-orientovaná, dokumentová, klíč-hodnota.

Hlavní aspekty realizace

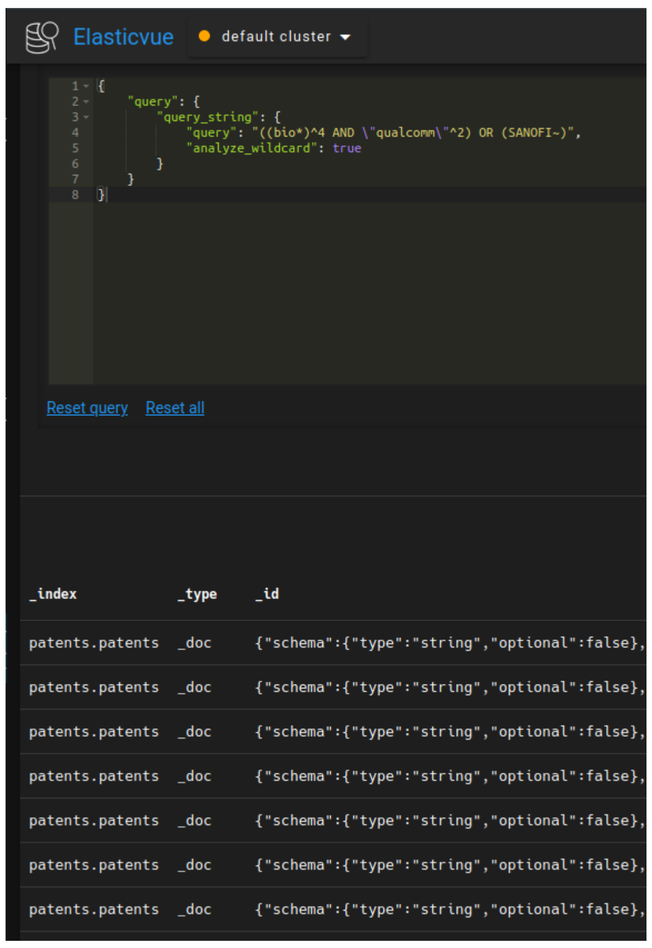
Výběr dat závisel na několika podmínkách: data musela obsahovat 4 povinné atributy (datum, identifikátor, titulek a autor), data lze stáhnout bez poplatků a data musí obsahovat pouze záznamy o užitných patentech. U dat byly dále sledovány i nepovinné atributy jako například: obor, typ, abstrakt, klíčová slova.

Validní data byla následně dále filtrována podle data přihlášení / publikace patentu. Všechny patenty před rokem 2000 byly vyfiltrovány, společně s těmi, kteří neobsahovali hodnotu u povinného atributu (elementu).

Struktura patentů se u každé země lišila, v některých případech (například Francie) byla struktura patentu změněna 3krát během 4 let. Vzhledem k různosti struktur dat musela být použita pro uložení dat databáze dokumentů, která umožňuje skladovat data s rozdílnýma strukturama.

V databázi dokumentů je možné data prohledávat full-textově, ale lepší výsledek přináší spojení databáze a speciálních vyhledávacích nástrojů, i za cenu větších nároků na paměť.

U patentů je také vhodné sledovat například v jaké zemi byly registrovány, v jakém jazyce jsou napsány a jakou mají klasifikaci. Vzhledem k rozdílným strukturám patentů a různým způsobům uložení specifických hodnot je dokumentová databáze nevyhovující, proto byla implementována druhá databáze, konkrétně relační databáze.



Ukázka full-textového vyhledávání v databázi patentů

Dosažené výsledky

Celkem bylo prozkoumáno 51 národních patentových zdrojů, ze kterých 33 zdrojů neposkytovalo žádná patentová data, 4 zdroje poskytovaly data za peníze, 4 zdroje poskytovaly nepoužitelná data, a zbývajících 10 zdrojů poskytovalo data zdarma s validními daty.

Celkový počet stažených patentů se blížil ke 3 milionům, při filtrování bylo nalezeno přibližně 1 milion patentů s neúplnými hodnotami.

Lokální úložiště sestává z 2 databází a 1 vyhledávacího nástroje. Jako nejlepší databáze pro vytěžování statistik byla použita relační databáze MySQL, pro full-textové vyhledávání byla použita databáze MongoDB s vyhledávacím nástrojem Elasticsearch.

Závěr

Vzhledem k již existující diplomové práci byly zkoumány pouze národní zdroje patentových dat, kde pouze malá část z nich poskytovala svá data zdarma. Pro tyto data bylo vytvořeno lokální úložiště sestávající z databáze MySQL, které umožňuje efektivně získávat statistiky z patentů, a databáze MongoDB, ve které jsou uložena veškerá patentová data. Nástroj Elasticsearch byl použit pro vyhledávání v databázi MongoDB.

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI