



Tvorba rozsáhlých úložišť patentových dat

Vojtěch Danišík¹

1 Úvod

Patenty jsou v dnešní době nedílnou součástí zákonné ochrany nových technologií a vynálezů v průmyslu a jejich odvětví. Vlastník takového patentu má výhradní právo využívat vynález a poskytovat souhlas o využití patentu jiným stranám. Patenty jsou registrovány a spravovány jak v rámci jednotlivých zemí (národní patentové instituce), tak i mezinárodně (například Evropský patentový úřad EPO). Tyto instituce veřejně poskytují informace o registrovaných patentech pomocí vyhledávačů na jejich webových stránkách, a v některých případech dokonce i plný export jejich databáze. Pro vědce a vynálezce jsou tyto vyhledávače velice důležité právě proto, aby si mohli zkontrolovat, zda jejich vynález je jedinečný, nebo už ve světě existuje v lehce upravené podobě a je patentován.

2 Výběr patentů

Celosvětové patentové instituce nebyly prostudovány z důvodu již existující diplomové práce (Baratta, S. (2019)), která byla zaměřená právě na tyto celosvětové instituce. Tato diplomová práce rozšiřuje původní diplomovou práci ve směru národních patentujících institucí a jejich dat.

Při výběru patentů byly stanoveny čtyři podmínky, které museli být splněny: **Dostupnost** (Patenty musí být dostupné z online stránek / databází a bez poplatků), **Datum** (Patentová přihláška nebo publikace patentu musí být podána alespoň v roce 2000, všechny ostatní patenty budou vyfiltrovány), **Atributy** (Všechny patenty musí obsahovat povinné atributy (titulek, datum, autor, identifikátor)) a Alternativy jako Užitný a Průmyslový vzor (Návrhový patent) nebudou brány v potaz.

Průzkum národních zdrojů zahrnoval celkem 51 národních patentujících institucí, ze kterých pouze 10 poskytovalo svá data zdarma a splňovala všechny podmínky. V tabulce č. 4.7 lze vidět souhrn výsledků.

Popis	Počet	Poměr
Nedostupné	33	64,70 %
Nepoužitelné	4	7,85 %
Za peníze	4	7,85 %
Použitelné	10	19,60 %

Tabulka 1: Tabulka příkazů použitých při tvorbě abstraktu

¹ student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, specializace Softwarový inženýr, e-mail: danisik@students.zcu.cz

V tabulce č. 2 lze vidět výsledný počet patentů všech zemí, jejichž patentující instituce poskytují svá data zdarma. Z celkových dat byla vyfiltrována přibližně 1/3 patentů, protože neobsahovaly hodnoty u všech povinných atributů (titulek, datum, identifikátor a autor).

Země	Počet patentů před filtrací	Filtrace	Počet patentů po filtraci
Velká Británie	88 032	1	88 031
Francie	746 899	474 850	272 049
Israel	116 380	326	116 054
Itálie	7 624	530	7 094
Kanada	936 464	119 696	816 768
Litva	869	0	869
Peru	1 805	0	1 805
Portugalsko	69	0	69
Rusko	614 256	223	614 033
Španělsko	381 713	308 064	73 649
Souhrn	2 894 111	903 690	1 990 421

Tabulka 2: Výsledný počet patentů po filtraci.

3 Úložiště

Dokumentová databáze bude použita jako primární databáze, protože umožňuje ukládat nestrukturované dokumenty velice efektivně. Zároveň podporuje vkládání dokumentů ve více formátech, což v případě mnoha národních zdrojů, kdy každý zdroj ukládá dat v jiném formátu, je velice vhodná vlastnost. Její nevýhoda, kontrola konzistence, může být odstraněna pomocí jednoduché aplikace, která bude kontrolovat výskyt patentu v databázi podle jeho identifikátoru (ID). Relační databáze bude použita jako sekundární databáze pro vytěžování, zaměřená hlavně na tvorbu statistik. Relační databáze je vhodnou volbou pro získávání statistik, protože vyhledávání je velice rychlé a snadné. Lze zmínit i nevýhody jako třeba problém s údržbou nebo potřeba velkého množství paměti. Tyto nevýhody ale nehrají velkou roli v případě jednotek až desítek milionů záznamů.

4 Závěr

Jako existující řešení databáze dokumentů byla vybrána komunitní verze databáze **MongoDB** spolu s full-textovým vyhledávačem **Elasticsearch**. Propojení mezi MongoDB a Elasticsearch je zařízeno pomocí **Apache Kafka**, což je distribuované úložiště událostí a platforma pro zpracování datových proudů. Jako existující řešení relační databáze byla vybrána komunitní verze databáze **MySQL**. Instalace a nastavení veškerých nástrojů a aplikací pro správu databází je dosaženo pomocí nástroje **Docker**.

Literatura

BARATTA, S. Creation of Data Sources for Bibliometric Analysis. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, 2019.