KnowING IPR:
projekt podpory inovací znalostními prostředky

Karel Ježek, Dalibor Fiala, Martin Dostal, Štěpán Baratta, Pavel Herout, Ladislav Pešička, Markéta Včalová, Pavel Král, Michal Nykl, Ladislav Lenc

Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni
Univerzitní 8, 306 14 Plzeň
{jezek_ka, dalfia, madostal, stepanb, herout, pesicka, mkucova, pkral, nyklm, llenc}@kiv.zcu.cz

Abstrakt. V příspěvku jsou popsány cíle a současný stav řešení české části mezinárodního projektu, který sdružuje země podunajského regionu. Za podpory Evropské unie a jejího programu pro regionální rozvoj Interreg má projekt za úkol zdokonalit podmínky pro inovace. Na základě zmapování současného stavu využívání práv duševního vlastnictví (IPR), zejména z průmyslové oblasti, klade si projekt za cíl vytvořit nadstavbu k současným bázím dat, kterými disponují patentové úřady, univerzitní knihovny, technické časopisy a obdobné datové zdroje zúčastněných zemí. Tato nadstavba propojí jednotlivé databáze a umožní získávat informace o technických řešeních uživatelsky příjemnějším způsobem a v komplexnější formě.

Klíčová slova: duševní vlastnictví, přenos technologií, patenty, průmyslové vzory

1 Úvod

Cílem projektu je zlepšení rámcových podmínek inovačních procesů v zemích patřících do povodí Dunaje. Projekt je součástí mezinárodního programu INTERREG DANUBE, jehož programové území pokrývá čtrnáct zemí střední a jihovýchodní Evropy od jižního Německa po Bulharsko. Konkrétně se jedná o část programu „Innovative and socially responsible Danube region“, která má rozpočet téměř 76 000 000 €. Název projektu KnowING IPR je akronymem pro „Fostering Innovation in the Danube Region through Knowledge Engineering and Intelectual Property Rights (IPR) Management“. Projekt je důsledkem přiznání významu IPR pro ekonomický rozvoj, prezentovaný např. v [2], [4] nebo [5].

KnowING IPR poskytne platformu volně přístupných prostředků pro analýzu IPR a návody pro zlepšení a harmonizaci IPR podmínek v regionu Dunaj. Zajistí bohatší informační zdroje o stávajících inovacích, výsledcích výzkumu, patentech a IPR znalostech a zlepší podmínky pro komercionalizaci výsledků výzkumu a transfer technologií (TT). Jedná se o pionýrský počin: využití pokročilých technologií znalostního inženýrství v oblasti IPR. Umožní se tím sdílené využívání existujících inovací a

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 1-4.

zlepšení příležitostí pro spolupráci založenou na IPR (předávání a přebírání znalostí a licencí). Projekt otevře dosud přiškrcený trh IPR a podnítí investice do inovací. Věříme, že také vytvoří konkurenční výhodu pro menší podniky, vysoké školy a výzkumné instituce dunajského regionu. Konečným produktem projektu bude volně přístupný „znalostní hub“, sdružující jednotlivé databáze, umožňující jejich simultánní dotazování a poskytující pokročilé analytické funkce. Tento znalostní systém bude pracovat v on-line režimu a bude umět zodpovědět dotazy typu „Kolik stojí patentová přihláška v Rumunsku?“, „Kdy vyprší platnost patentu X v Maďarsku?“ apod.

2 Způsob řešení

Do projektu se zapojilo celkem šestnáct partnerů z univerzit, výzkumných institucí a vládních úřadů ze třinácti různých zemí. Řešení bylo zahájeno 1. 7. 2018 a bude probíhat do 30. 6. 2021 s rozpočtem 2 149 800 €, z něhož 270 000 € připadne na ZČU v Plzni.

Celý projekt je rozfázován do několika etap, z nichž ta současná končí říjnem 2019. Tato část má za úkol popsat v jednotlivých zemích aktuální stav IPR, TT, zejména pak datové kolekce a jejich současnou i potenciální použitelnost v projektu. Obdobným způsobem je prováděno ohodnocení mezinárodních databází a významných databází institucí z třetích zemí, jako je např. PATSTAT (European Patent Office), PatentScope (World Intellectual Property Organization), či USPTO (United States Patent and Trademark Office). Za relevantní jsou považovány i vědecké a firemní databáze, jako jsou ACM Digital Library, CiteSeer, DBLP, Google Scholar, Microsoft Academic Graph, PubMed, Scopus, SemanticScholar, Web of Science, apod.

Po etapě vyhodnocující vhodnost a dosažitelnost dat, následuje jejich akvizice. Touto činností je pověřena část nazvaná *Data Acquisition Module* (DAM). Data jsou ke stažení z webových stránek v různých formátech, nejčastěji ve formátu XML a JSON, ale také HTML, CSV, XSL, TXT, ZIP nebo PDF. Je proto nutné před ukládáním provést jejich konverzi a parsing do formátu JSON importovatelného do nerelační DB (viz níže). Významné patentové databáze přitom mají rozsah dat i přes 100 GB.

Další funkcí související se stažením a uložením dat je jejich aktualizace. Administrační modul musí v pravidelných intervalech kontrolovat URL stahovaných dat, pomocí časových razítek rozpoznat, zda neobsahují nová data a případně rovnou provést jejich stažení, či upozornit na potřebu manuálního stažení administrátora systému.

Vývoj a struktura akvizičního modulu jsou významně ovlivněny zvoleným databázovým systémem (DBMS), který obhospodařuje hlavní, tzv. zdrojovou databázi s údaji o patentech a článcích. Po zvážení předností i nedostatků tradičních relačních versus NoSQL databází byl jako primární DBMS vybrán MongoDB, který se již dříve osvědčil při zpracování nestrukturovaných dat ve znalostních aplikacích [1]. Zdrojová databáze, modelovaná v nerelační architektuře MongoDB bude obsahovat dva typy kolekcí: pro publikace a pro patenty.

Každá z kolekcí bude ukládat dokumenty z odlišných datových zdrojů, jejichž struktura se sice bude lišit, ale části jako osoby, firmy apod. budou součástí obou. Jelikož MongoDB je dokumentografická databáze, jejími základními prvky jsou do-

kumenty, které mohou mít různou strukturu. Stažená a extrahovaná data, případně převedená do JSON formátu, jsou částečně restrukturována před uložením do cílové databáze systému MongoDB. DAM provádí také kontrolu a odstranění duplicit v záznamech a určení klíčových slov pro každý záznam.

Dva základní typy kolekcí znamenají také dva způsoby řešení cílové databáze:

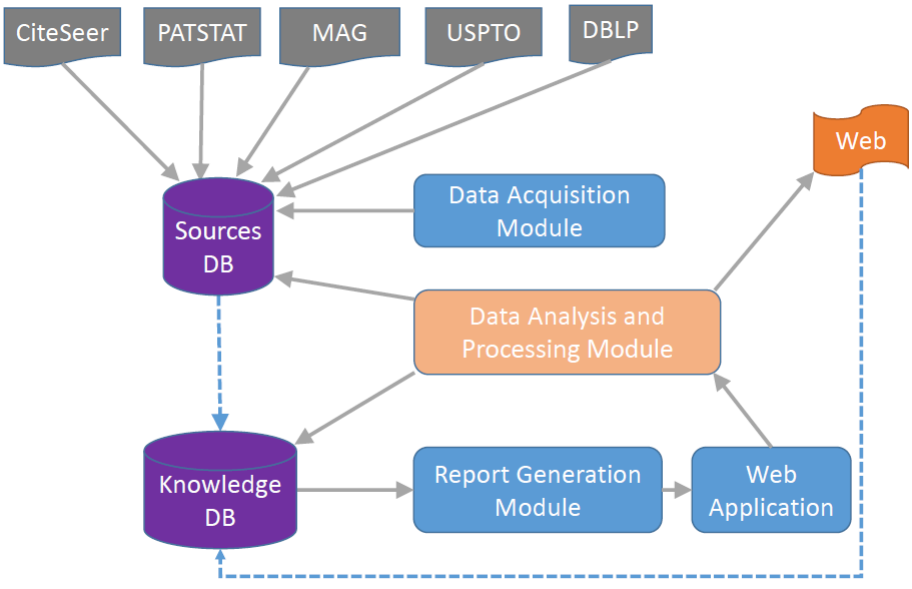
- Separátní MongoDB kolekce pro publikace a pro patenty pro každý datový zdroj.
- Jedinou společnou cílovou kolekci pro každý typ dat.

Zvolen byl druhý způsob, pro jednodušší vkládání i dotazování, pracující pouze se dvěma globálními kolekcemi: *patents* a *publications*.

S daty uloženými ve zdrojové databázi nadále pracuje *Data Analysis and Processing Module* (DAPM). Oproti databázově orientovanému DAM, je DAPM pověřen znalostními funkcemi. Vyhodnocuje a zpracovává IPR dotazy uživatelů a ukládá odezvy na dotazy uživatelů prostřednictvím *Report Generation Module* (RGM) a *Core Communication Controller* (CCC) do znalostní databáze (*Knowledge Database* - KD). KD je SQL databáze realizovaná na bázi systému MariaDB. DAMP, RGM a CCC tvoří tzv. *Knowledge Generation Core* (KGC).

Po zadání dotazu uživatelem přes webovou aplikaci probíhají následující aktivity:

1. Předání dotazu do KGC, jeho lemmatizace a extrakce klíčových slov.
2. DAPM ověří, zda dotaz byl již zadán v minulosti.
3. Pokud byl, vybere se odpověď z KD a je předána uživateli.
4. Pokud nebyl, je dotázána zdrojová (NoSQL) databáze a případně web. Ze seznamu navrácených výsledků je vytvořena odpověď v JSON formě a je zaslána tazateli.
5. Nově získaná odpověď se spolu s dotazem zaznamená do KD (viz Obr. 1).



Obr. 1: Blokový diagram se schématem projektových modulů a databází

3 Současný stav projektu

Řešitelé projektu byli poměrně přesně instruováni, jak postupovat při analýze situace IPR v jednotlivých zemích. Instrukce zahrnovaly anketní otázky, na které odpovídali experti z akademické, průmyslové i politické sféry. Pro vyhodnocení anketní části byl uspořádán workshop s účastí expertů, který formuloval závěry hodnocení stavu IPR v ČR a jeho vliv na výzkum a vývoj. Provedená SWOT analýza věnovala zvláštní pozornost těm slabým stránkám využití IPR, které jsou ovlivnitelné realizací projektu KnowING IPR. Jedná se zejména o:

- nedokonalé rešeršní patentové služby,
- nákladné právní poradenství,
- rozmanitost informačních zdrojů a jejich obtížnou přístupnost.

Z těchto důvodů bylo rozhodnuto v akviziční fázi stáhnout data z národních patentových databází alespoň z části dostupných v anglickém jazyce. Na této úloze se podílí každý z národních týmů. My jsme se zaměřili na získání a propojení dat z české a evropské databáze patentů.

4 Závěr

Propojení databází je pouze prvým krokem k vytvoření znalostního systému, jenž dataminingové komunitě zpřístupní velké množství heterogenních dat z oblasti patentů. Báze dokumentů i znalostí se však musí průběžně doplňovat a měnit. Vytvoření dynamického znalostního systému, který bude obdobně jako v [3] pracovat v režimu „Never Ending Learning“ je proto naším dalším cílem.

Poděkování: Tento příspěvek vznikl s podporou programu Interreg Danube, projektu „KnowING IPR: Fostering Innovation in the Danube Region Through Knowledge Engineering and IPR Management“ (číslo projektu DTP2-076-1.1).

Literatura

1. Jabbari S., Stoffel K.: Ontology Extraction from MongoDB Using Formal Concept Analysis, In: Proc. 2nd International Conference on Knowledge Engineering and Applications (ICKEA 2017), London, UK, 21 - 23 October 2017, 178-182.
2. Kogan, L., Papanikolaou, D., Stoffman, A.S.N.: Technological Innovation, Resource Allocation, and Growth. The Quarterly Journal of Economics 132 (2017), 665–712.
3. Mitchell, T., Cohen, W., Hruschka, E. et al.: Never-Ending Learning. Communications of the ACM 61 (2018), 103-115.
4. Tolstaya, A.M., Suslina, I.V., Tolstaya, P.M.: The role of patent and non-patent databases in patent research in universities. AIP Conference Proceedings 1797 (2017), no. 020017.
5. Van Raan, A.F.J.: Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research. Journal of Data and Information Science 2 (2017), 13-50.