Payola - Framework pro vizualizaci grafových dat na webu

Vedoucí:

- Martin Nečaský, Jakub Klímek

Řešitelé:

- Helmich Jiří, Heřmánek Ondřej, Kudláček Ondřej, Široký Jan, Váša Kryštof

Motivace

Grafová data označují data tvořená entitami a vztahy mezi nimi. Jsou charakteristická tím, že nemají žádné a nebo jen volně dodržované schéma. Tím se výrazně liší např. od striktně strukturovaných dat v relačních databázích. Příkladem jsou např. RDF data dostupná na webu v podobě tzv. Linked Data[1] či data dostupná na sociálních sítí typu Facebook či LinkedIn. Také řada firem zpracovává svá interní data, která mají podobu grafových dat.

Možnost propojovat, analyzovat a vizualizovat grafová data v uživatelsky přívětivé podobě je dnes proto velmi žádaná. Existující nástroje jako např. http://opendata.socrata.com, http://www.gooddata.com, http://www.google.com/publicdata/, apod. však umožňují pouze práci s tabulkovými daty. Naproti tomu mají řadu zajímavých funkčností - např. uživatelé mají možnost vytvářet svoje vlastní analýzy dat a sdílet či prodávat svá data a analýzy mezi sebou.

Některé nástroje pro vizualizaci či analýzu grafových dat existují, ale ty jsou striktně fixované jen na konkrétní datovou sadu, např. obchodní rejstřík: http://obchodni-rejstrik.podnikani.cz/applet/

Cíl projektu

Cílem projektu je navrhnout a implementovat prototyp SW systému v podobě webové aplikace, která umožní svým uživatelům pracovat s libovolnými grafovými daty - tj. data importovat, analyzovat a vizualizovat.

Funkční požadavky

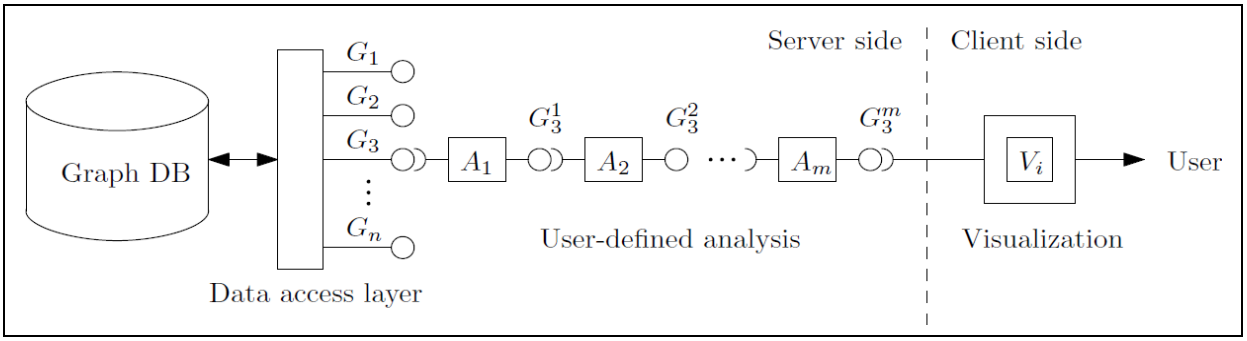
Hlavní funkční požadavky na systém jsou následující:

- Uživatel vlastní a spravuje svůj privátní datový prostor.
- Uživatel může přidávat do svého datového prostoru nová grafová data několika způsoby:
    - upload lokálních dat ze svého disku

[1] http://linkeddata.org/

    - import datové sady ze zadaného URL (např. viz katalog zdrojů grafových dat na webu v podobě Linked Data: http://linkeddata.org/)
- Uživatel může sdílet svá data s vybranými uživateli
- Uživatel může prostřednictvím uživatelského rozhraní definovat analýzu grafových dat jako sekvenci zvolených analytických pluginů následujících typů:
    - ontologický filtr (vstupem je obecný graf, výstupem je graf odvozený ze vstupního pomocí ontologie dodatené jako parametr)
    - grafový algoritmus (vstupem je obecný graf, výstupem je graf nalezený daným grafovým algoritmem, např. nejkratší cesta)
    - grafová projekce (obdoba projekce nad relační tabulkou v grafovém modelu)
    - grafová selekce (obdoba selekce nad relační tabulkou v grafovém modelu)
    - transformace grafu (vstupem je obecný graf, výstupem je výsledek vyhodnocení dodaného dotazu v jazyce SPARQL nad vstupním grafem)
- Výstupem analýzy bude vždy opět graf. Uživatel může výstup analýzy vizualizovat pomocí:
    - generické vizualizační metody nebo
    - pomocí zvoleného vizualizačního pluginu specifického pro vybranou ontologii
- Uživatel může s ostatními uživateli sdílet svoje specifikace analýz i vizualizace výstupů.

Architektura aplikace je znázorněna na obrázku 1. Skládá se z databáze, serverové a klientské části. Serverová část je realizována frameworkem, který umožňuje přístup do databáze ke grafovým datům a spouští tzv. analytické workflows. Analytický workflow je sekvence na sebe navazujících analytických pluginů, jejichž typy byly uvedeny výše. Klientská část přijímá výsledky analýzy ze serverové části a vizualizuje je uživatelem zvoleným způsobem. Uživatel v klientské části také nastavuje analytické workflows.



Obrázek 1: Architektura aplikace. Datová vrstva poskytuje přístup k datovým prostorům jednotlivých uživatelů. Uživatel může nadefinovat posloupnost analytických pluginů, které jsou postupně na jeho graf aplikovány. Výsledek je zobrazen visualizérem s použitím vybraného visualizačního pluginu.

Nefunkční požadavky

Z důvodů přenositelnosti, interaktivity a možností škálovatelnosti, které moderní technologie nabízí, bude prototyp systému implementován ve formě webové aplikace. Pro lepší uživatelský komfort bude využita technologie AJAX, díky které je možné implementovat některé funkce uživatelského rozhraní tak, jak jsou uživatelé zvyklí z desktopových aplikací. Pro zachování co nejvíce transparentního kódu a maximalizaci jeho znovupoužití bude využita technologie pro překlad kódu z programovacího jazyka serverové části aplikace do kódu klientského programovacího jazyka (JavaScript). Pokud bude třeba, bude tento mechanismus doplněn o jednoduchou RPC bránu, která usnadní komunikaci mezi klientem a serverovou částí aplikace.

Vzhledem k tomu, že systém bude pracovat s grafovými daty, bude součástí projektu také výběr vhodné grafové databáze, nad kterou je možné navrhovaný systém implementovat. Tato databáze bude sloužit jako úložiště dat a měla by také poskytovat rozhraní pro dotazování nad těmito daty. Pokud bude třeba, budou navrženy postupy a opatření, jak optimalizovat práci s touto databází za účelem rychlejšího získání odpovědí na položené dotazy. S tím souvisí také implementace datové vrstvy, která odstíní business logiku aplikace od práce s konkrétní použitou grafovou databází.

Výstup analytických algoritmů také není dopředu znám, vzhledem k tomu, že bude zcela záviset na preferencích uživatelů. Při návrhu architektury systému budeme předpokládat, že analytické algoritmy budou často hledat různé podgrafy analyzovaných grafů, případně nad nimi budou počítat libovolné charakteristiky. Systém musí poskytnout dostatečně flexibilní rozhraní na integraci takových algoritmů, nejlépe ve formě plug-inů, které se na sebe dají řetězit.

Zásadní bude vyřešení komunikace mezi klientskou a serverovou částí aplikace. Analyzovaná grafová data totiž mohou být libovolně velká a uživatel prostřednictvím klientské části aplikace vždy vidí jen určitý výsek. V rámci projektu tedy budou navrženy vhodné optimalizace založené na předpočítávání a kešování výsledků na straně serveru i klienta. Dále bude nutné navrhnout vhodný komunikační protokol pro efektivní výměnu grafových dat a uživatelských požadavků na vizualizovanou část dat mezi klientem a serverem.

Jako ukázku použití navrženého systému bude implementována Databáze korupčních případů jejíž součástí budou data z obchodního rejstříku ČR (anonymizovaná z důvodu zákona o ochraně osobních údajů). Základní verze bude vycházet z požadavků Nadačního fondu proti korupci (http://www.nfpk.cz). Rozšíření systému pak bude demonstrovat práci s navrženou aplikací a integrovat některé testovací analytické plug-iny.

Další požadavky na projektový tým

- k projektu vznikne hned od počátku webová stránka, kde postupně bude vznikat informační, uživatelská a programátorská dokumentace k vytvářenému software včetně možnosti jeho stažení
- veškerá vytvářená dokumentace bude v anglickém jazyce

Předpokládaný průběh práce

V projektu bude implementován prototyp specifikovaného SW systému. Projekt je dimenzován na práci 5ti členného týmu v průběhu 9ti měsíců. Práce budou probíhat dle následujícího schématu:

1. Analýza existujících implementací a přístupů [M1-M2]
2. Podrobná specifikace konkrétních funkcí systému, architektury a rozhraní mezi jednotlivými moduly [M2-M3]
3. Implementace prototypu [M3-M8]
4. Testy na reálných datech, ladění [M8-M9]
5. Dokumentace (programátorská, uživatelská, instalační) [M7-M9]