Distance search

Ján Garaj
Fakulta informatiky a informačných technológií
Slovenská technická univerzita
Školský rok: 2008/09

Popis problému a motivácia

Vyhľadávanie podľa vzdialenosti hľadá relevanciu k dotazovanému objektu (referenčnému bodu) podľa vzdialenosti. Vzdialenosť je možné špecifikovať rôzne, napr. bitová vzdialenosť, vektorová vzdialenosť, fyzická vzdialenosť, komunikačná vzdialenosť. Riešený problém v tomto projekte rieši vyhľadávanie nad doménou firiem a spoločností uverejnených na verejne dostupných stránkach slovenského portálu www.zoznam.sk podľa geografických umiestnení (vzdialeností) sídel spoločností. Každá spoločnosť má na zoznam.sk uverejnené svoje sídlo, pričom obec sídla spoločnosti je považovaná za základ pri vyhľadávaní. Každá obec má presne stanovenú svoju zemepisnú šírku - latitude a zemepisnú výšku longtitude. Z týchto súradníc označujúcich zemepisnú polohu obce je možné následne vypočítať ich vzdialenosť. Reálny vzorec na výpočet vzdialenosti musí zohľadňovať aj zakrivenie zemegule, avšak vzhľadom, že daná doména firiem a spoločností uverejnených na www.zoznam.sk je lokalizovaná na relatívne malé územie je možné v prípadných výpočtoch zakrivenie zemegule zanedbať. Potrebné je si však aj uvedomiť, že vzdialenosť medzi dvoma bodmi (obcami) je ich vzdušná vzdialenosť a ich reálna vzdialenosť napríklad po cestných komunikáciách je vždy reálne dlhšia. Vyhľadávania podľa vzdialenosti je praktické pri vyhľádávaní v zadanej lokalite, resp pri hľadaní podobných objektov umiestnených v blízkosti referenčného bodu. Napríklad nájdi mi všetky betonárne na tomto definovanom území. Alebo nájdi mi najbližšie obuvníctvo od tohto obuvníctva, keďže v tomto obuvníctve nemajú také topánky ako potrebujem, možno ich budú mať v inom blízkom obuvníctve. Praktické využitie teda je možné nájsť vyhľadávaní blízkych ekvivalentov referenčných objektov (služieb). Minimalizuje sa tým prejdená vzdialenosť, prípadne prepravná vzdialenosť, keďže pri použití klasického vyhľadávania sa geografická poloha nezohľadňuje vo výsledkoch vyhľadávania. Z uvedého vyplýva hlavná výhoda hlavne v minimalizácii nákladov a skracovaní potrebných časov.

Prehľad súčasných riešení:

Napriek tomu, že sa môže jednať o zaujímavé vylepšenie, bežne slovenské vyhľadávače neimplementujú priamo distance search, avšak zvyčajne iba obmedzujú vyhľadávanie na kraje, okresy prípadne jednotlivé obce (napríklad centrum.sk).

Vyhľadávaním, ktoré je možné označit za distance search disponuje spoločnosť Google vo svojom produkte Google Maps, cez funkcionalitu "Search near by". Po zadaní klúčového slova sa zobrazia výsledky, ktoré zodpovedajú kľúčovému slovu a sú v predefinovanej vzdialenosti. Vzdialenosť je možné zmeniť v niekoľkých krokoch od 800m do 150km. Nevýhoda tohto vyhľadávania, je že vyhľadáva iba v objektoch POI (Point of interest), ktoré vyznačili a vytvorili používatelia systému maps.google.com.

Na veľkej väčšine webových stránok je distance search ponímaný vo forme hľadania najkratšej (najrýchlejšej) cesty medzi dvoma mestami. Toto ponatie nie je však úplne správne, nakoľko v tomto prípade sa jedna o hľadanie najratšej prípadne najlacnejšej cesty medzi dvoma bodmi v grafe. Na tieto problémy sú vhodné algoritmy na prehľadávanie grafu. Tieto metódy je možné aj využiť v distance search pri určovaní reálnej cestnej vzdialenosti objektov.

Popis riešenia

Návrh riešenia vychádza z použitia ontológie ontológie geografických údajov slovenska v OWL [1]. Táto ontológia obsahuje aj záznam o zemepisnej šírke a dĺžke, ktoré sa dajú využiť pri vyhľádávani distance search. Problémom je že ontológia neobsahuje žiadne dáta o spoločnostiach, takže túto ontológiu je potrebné rozšíriť aj o spoločnosti, ktoré sa vyextrahujú z portálu zoznam.sk

Extrakcia dát a tvorba ontológie spoločností slovenska

Postup pri extrakcii dát z portálu zoznam.sk:

1.) Vytvorenie úplného obrazu celého portálu pomocou nástroja wget – trvanie približne 2 dni

2.) Z vytvorenej mirror databázovej štruktúry sa vytvoril zoznam html súborov, ktoré popisovali jednu spoločnosť – 29175 html súborov

3.) Vytvorenie regulárnych výrazov na výber názvu spoločnosti, kategórií do ktorých je zaradená spoločnost, emailov, faxov a web stránok spoločnosti, adresa, obrat, počet zamestnancov a pod. Niektoré dáta sa aj upravovali na jednotný formát. V prípade, že sa z html súboru nepodarilo vybrať niektoré dáta nutné na vytvorenie

spoločnosti bola celá extrakcia dát z daného súboru zrušená. Vyextrahované dáta sa ukladali do relačnej databázy. Počet vyextrahovaných spoločností 29109.

4.) Práca s nástrojom Protege - rozšírenie ontológie o triedu Company, ktorá má reláciu HasRegion, ktorá určuje sídlo spoločnosti, pričom je to vlastne referencia na už existujúcu inštanciu v ontológii geografických údajov slovenska. Názov spoločnosti sa ukladá do novej dátovej vlastnosti triedy Company – title.

5.) Vytvorenie 3 testovacích inštancií triedy Company a export ontológie do formátu OWL. Následné štúdium formátu uložených inštancií umožnilo vytvoriť php skript, ktorý zapisoval existujúce spoločnosti v relačnej databáze do OWL (XML) formátu. Tento skript sa taktiež pokúšal nájsť inštanciu sídla spoločosti v existujúcej ontológii. Problém, ktorý nastal bola skutočnosť, že názvy obcí v ontológii Slovenska boli bez diakritiky, takže skript sa pokúšal hľadať aj názvy obcí bez diakritiky. Nie vždy však bolo hľadanie úspešné, nakoľko napríklad sídlo Košice nebolo v ontológii Slovenka, avšak mestké časti Košíc ontológia obsahovala. Z tohto dôvodu sa nie vždy podarilo vytvoriť OWL inštanciu. Počet úspešne vyexportovaných inštancií spoločnosti bol 20568.

6.) Aby sa bolo možné dotazovať nad ontológiou je potrebné novovytvorenú ontológiou uložit. Ako úložisko bol zvolený ontologický repozitár Sesame nad ktorým pracujú a cez ktorý sa dotazujú samotne vyvíjané aplikácie pre distance search.

Opis programu

Pôvodný návrh aplikácie počítal s hľadaním v určitom okruhu od 1 referenčného bodu. Tento pôvodný návrh však musel byť pozmenený na vyhľadávanie v obdĺžniku určenom 2 geografickými bodmi. Problém vzniká totiž v použitej technológii Sesame. Dotazovací jazyk SPARQL v Sesame, ktorým sa dotazuje nad ontológiami, totiž nemá funkcie, ktoré by umožnovali vo vyhľadavacích kriteriách definovať alebo vypočítavať geografickú vzdialenosť dvoch bodov [2]. Štandard SPARQL umožňuje definovať si vlastné funkcie, avšak ontologický nástroj Sesame neobsahuje túto funkcionalitu. V prípade, žeby umožnoval, napríklad ako nástroj ARQ je možné definovať funkciu, ktorá umožňuje vyhladávať aj podľa sférickej vzdialenosti 2 bodov [3].

Vytvorený program je vlastne klient, ktorý sa dopytuje prostredníctvom jazyka SPARQL nad Sesamom. Vstupom sú súradnice dvoch bodov, podľa ktorých sa dotaz vymedzí. Z dôvodu rýchlosti reakcie celého systému dotaz obsahuje aj limit na 1000 výsledkov.

Výsledný dotaz:

```
SELECT ?name ?title ?longitude ?latitude
WHERE {
?x location:name ?name .
?x location:hasRegion ?b.
?b location:title ?title .
?b location:latitude ?latitude .
?b location:longitude ?longitude .
FILTER (?latitude >= LA1).
FILTER (?latitude <= LA2).
FILTER (?longitude >= LO1).
FILTER (?longitude <= LO2).
}
ORDER BY ASC(?name)
LIMIT 1000
```

Použitie softvéru

Web GUI rozhranie je implementované nad maps.google.com a jeho API rozhraním, pomocou ktorého sa získavajú súradnice na mape. Takto získané súradnice sa pomocou AJAXu pošlú na vykonný php skript, ktorý zavolá skompilovaného java klienta a ten mu vráti naformátované výsledky. Cesty v samotnej aplikácii sú prednastavené na adresu http://localhost/vinf, takže prípadné pri inom umiestnení je potrebné nastavanie systému (napr. licenčný klúč na maps.google.com) upraviť.

Vyhodnotenie

Subjektívne

Sesame ako úložisko ontológií nie je moc vhodný v súčasnom stave na prácu s geografickými dátami. Vytvorený softvér prehľadával iba reláciu HasRegion, avšak v reálnom nasadení by sa ešte prehľadávali napríklad aj relácie príslušnosti ku zadanej kategórii prípadne podobnosť názvu spoločnosti so zadaným textom. Predpokladám, že rýchlosť reakcie Sesame by v tomto smere ešte klesla pod hranicu, ktorú by používateľ vnímal ako kritickú. Na strane druhej semantický web a ontológie sa považujú za budúcnosť internetu, takže je očakávané zrýchlenie aplikácií na prácu s ontológiami –

Sesame, na takú úroveň, aby mohli konkurovať relačným databázam.

Objektívne

Objektívne vyhodnotenie (Recall, Precision) nebolo možné realizovať v akceptovateľnej podobe, nakoľko priame vyhľadávanie spoločností neumožňuje aplikácia maps.google.com. Približné porovanie bolo zrealizované cez funkciu „"Search near by", pričom klúčové slovo „pizza“ bolo zadané v strede obce Ružomberok. Maps.google.com vrátil vo výsledku 4 pizzérie v tomto meste. Implementovaná aplikácia po vyznačení oblasti vrátila približne 900 výsledkov spoločností sídliacich vo vyznačenej oblasti Ružomberka. Medzi nájdenými spoločnosťami však bola iba jedna pizzeria. Vyhodnotenie je teda silne závislé na počte POI v maps.google.com a od počtu registrácií na zoznam.sk, čo sú silne nezávislé entity.

Použitá literatúra

[1] Ontológia geografických údajov slovenska v OWL
http://laclavik.net/projects/kwfgrid/location_data_slovakia.owl
[2] http://en.wikipedia.org/wiki/Great-circle_distance
[3] Gregory Todd Williams: Extensible SPARQL Functions With Embedded Javascript
[4] http://www.scribd.com/doc/2569355/Geo-Distance-Search-with-MySQL