Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: INFORMAČNÉ SYSTÉMY

Bc. Štefan Dlugolinský

VYHĽADÁVANIE INFORMÁCIÍ NA WEBE PODĽA VZDIALENOSTI

Diplomový projekt I

Vedúci diplomového projektu: RNDr. Michal Láclavík PhD.

máj, 2008

Prehlásenie

Čestne prehlasujem, že som diplomový projekt vypracoval samostatne s použitím uvedenej literatúry.

Bratislava, máj 2008

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

ZADANIE DIPLOMOVÉHO PROJEKTU

Meno študenta: **Bc. Štefan Dlugolinský**
Študijný odbor: INFORMAČNÉ SYSTÉMY
Študijný program: Informačné systémy
Názov projektu: **Vyhľadávanie informácií na webe podľa vzdialenosti**

Zadanie:

Fulltextové vyhľadávače určujú relevantnosť podľa obsahu dokumentu ako aj významu stránky, ktorý sa zisťuje pomocou odkazov medzi stránkami a algoritmov ako PageRank. Pri hľadaní služieb pomocou kľúčových slov môžeme uprednostniť iné utriedenie výsledkov vyhľadávania, napríklad podľa služieb ktoré sú k nám najbližšie. Cieľom je teda navrhnúť a implementovať algoritmy a systém na spracovanie a vyhľadávanie informácií na webe tak, aby sa výsledky fulltextového vyhľadávania zoradili podľa vzdialenosti od referenčného bodu, pričom sú zaujímavé iba tie web stránky a web sídla kde je možné identifikovať geografickú lokalitu na základe algoritmov z oblasti vyhľadávania informácii (information retrieval) a extrakcie informácií pričom je potrebné priradiť lokalitám zemepisné dĺžky a šírky. Príkladom v tejto oblasti sú Google Maps alebo YellowPages Distance Search, pričom navrhnutý systém by sa mal odlišovať tým že je do určitej miery schopný určiť lokalitu stránky pomocou detekcie objektov ako adresy, mestá, ulice alebo PSČ. Systém je možné overiť v prostredí slovenského webu, napríklad vyhľadávanie služieb v rámci Bratislavy alebo Slovenska.

Odporúčaná literatúra:

1. Cunningham, H. (2005). Information Extraction, Automatic. Encyclopedia of Language and Linguistics, 2nd Edition
2. Otis Gospodnetic, Erik Hatcher: Lucene In Action; Manning Publications, December 2004

Práca musí obsahovať:

- Anotáciu v slovenskom a anglickom jazyku
- Analýzu problému
- Opis riešenia
- Zhodnotenie
- Technickú dokumentáciu
- Zoznam použitej literatúry
- Elektronické médium obsahujúce vytvorený produkt spolu s dokumentáciou

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava, Ústav informatiky SAV, Bratislava
Vedúci projektu: RNDr. Michal Laclavík PhD.
Pedagogický vedúci: Ing. Ivan Kapustík

Termín odovzdania práce v letnom semestri: dňa 19. mája 2008
Termín odovzdania práce v zimnom semestri: dňa 11. decembra 2008

Bratislava, dňa 18. februára 2008

prof. Ing. Pavol Návrat, PhD.
riaditeľ ÚISI

Obsah

Zoznam obrázkov

Zoznam tabuliek

Slovník pojmov

CSV – CSV (Comma-separated values) je formát súboru určený na výmenu tabuľkových údajov. Údaje sú v súbore oddelené čiarkou, prípadne bodkočiarkou alebo tabulátorom.

JSON – JSON (JavaScript Object Notation) je odľahčený formát na výmenu dát. Pre človeka je ľahko čitateľný i zapisovateľný a pre stroj ľahko analyzovateľný a generovateľný.

KML – KML (Keyhole Markup Language) je to formát súboru určený na prenos zemepisných údajov medzi prehliadačmi Google Earth, Google Maps, a Google Maps pre mobilné zariadenia.

RDF – RDF (Resource Description Framework) je rámec na výmenu metadát medzi aplikáciami [17].

RDF/XML – XML syntax na zápis RDF grafov.

REST – REST (Representational State Transfer) je architektúra určená pre distribuované systémy s hypermediálnym obsahom [3].

RPC – RPC (Remote Procedure Call) je technika vytvárania distribuovaných klient-server aplikácií. Hlavným princípom je možnosť volania procedúr externých systémov [8].

SOAP – SOAP (Simple Object Access Protocol) je XML protokol určený na výmenu údajov medzi aplikáciami prostredníctvom HTTP protokolu [18].

URI – URI (Uniform Resource Identifier) je ucelená sekvencia znakov identifikujúca abstraktný alebo fyzický zdroj. Ide o štandard popísaný v RFC 3986 [16].

URL – URL (Uniform Resource Locator) je adresa dokumentu alebo iného zdroja vo svetovej pavučine.

xAL – xAL (eXtensible Address Language) je aplikačne nezávislý XML štandard na reprezentáciu adries [13].

Úvod

Internet sa stal mocným komunikačným nástrojom a obrovským zdrojom informácií, čoho príčinou je jeho neustále rozširovanie a zvyšovanie prenosových rýchlostí. Vďaka rôznym technológiám je dostupný prakticky všade na Zemi. Práve kvôli nespočetnému množstvu informácií, nie je vyhľadanie tých relevantných jednoduchou záležitosťou. Existuje však veľa kvalitných vyhľadávačov a katalógov stránok, ktoré nám uľahčujú nájdenie toho, čo práve potrebujeme. Vyhľadané výsledky sú týmito fulltextovými vyhľadávačmi zoraďované najčastejšie podľa relevantnosti alebo abecedy. Častokrát však potrebujeme nájdené výsledky usporiadať podľa vzdialenosti od nejakého referenčného bodu. Na to však potrebujeme vedieť geografickú polohu webovského dokumentu a samozrejme aj vzťažnej pozície. Určenie referenčného bodu je problém pre používateľa vyhľadávacej služby, ale o niečo väčší problém a to lokalizácia webovského dokumentu, je problém pre vyhľadávací systém. Presnejšie povedané pre jeho analyzátor webovských dokumentov. Niektoré vyhľadávače z časti riešia problém vyhľadávania podľa vzdialenosti. Vedia napríklad zobraziť nájdené výsledky z určitej lokality. Problém je však s doménou, v ktorej sa vyhľadáva. Tá zväčša zahŕňa iba používateľmi definovaný obsah, kde používateľ určí geografickú polohu danej webovskej stránky, alebo je obsah viazaný na katalóg stránok, v ktorom sa pri registrácii stránok zadávajú kontaktné údaje a adresy. Podľa adresy sa potom zistí geografická lokalita stránky. V konečnom dôsledku ide teda o manuálne určovanie geografickej polohy internetových stránok. Je tu preto priestor pre automatické zisťovanie zemepisnej polohy stránok podľa ich obsahu. To by podstatne rozšírilo doménu vyhľadávania.

Tento dokument je správa o riešení diplomového projektu s názvom „Vyhľadávanie informácií na webe podľa vzdialenosti“. Nachádza sa v ňom prehľad existujúcich systémov, nimi používaných technológií a prístupov práve v súvislosti s vyhľadávaním informácií na webe podľa vzdialenosti. Analýza jednotlivých systémov je zameraná na problém geokódovania, respektíve zisťovanie geografickej lokality z webovského dokumentu. Ďalej je v dokumente načrtnutá problematika extrakcie informácií a sú analyzované dostupné prostriedky, ktoré by sa dali na extrakciu informácií využiť pri riešení diplomového projektu.

Analýza problému

Vyhľadávanie na webe podľa vzdialenosti v sebe ukrýva tri hlavné problémy, ktoré treba riešiť. Je potrebné nejakým spôsobom určiť geografickú polohu webovského dokumentu. Tu treba dokument podrobiť analýze a ak je to možné, extrahovať z neho informácie týkajúce sa jeho geografickej polohy. Ďalším problémom je pomocou týchto informácií zistiť zemepisnú dĺžku a šírku vzťahujúcu sa na dokument. Ak máme zistenú geografickú polohu, môžeme dokument podľa nej indexovať. A to je tretí problém, ktorý súvisí s usporiadaním výsledkov vyhľadávania podľa vzdialenosti od referenčného bodu. Ako to urobiť? Chceme predsa rýchly vyhľadávač a pri vyhľadávaní meniť referenčný bod. Takisto chceme mať čo najväčšiu bázu dokumentov. Ak by sme napríklad dokumenty, čo sa týka geografickej polohy, indexovali iba podľa geografických súradníc, museli by sme pri akejkoľvek zmene referenčného bodu pri vyhľadávaní, nanovo počítať ich vzdialenosť od východzej polohy. Pri veľkom počte dokumentov v bázi by to systém značne spomalilo. Bolo by prakticky jedno, či by sme spomedzi „najbližších dokumentov“ vyhľadávali najrelevantnejšie, alebo spomedzi najrelevantnejších „najbližšie“. V oboch prípadoch by sme mohli mať veľmi veľa „blízkych“ alebo veľmi veľa relevantných dokumentov.

Na internete existuje hojný počet informačných systémov, ktoré majú niečo spoločné s vyhľadávaním na mapách. Navzájom sa líšia rôznymi vlastnosťami ako je presnosť, pokrytie, dostupnosť, spoplatenenie služieb a pod. Vo svete sú známe systémy ako Google Maps, YellowPages.com alebo Yahoo! Local Maps. Na Slovensku je známy portál Kompas.sk. V nasledujúcich kapitolách si predstavíme spomínané systémy so službami, ktoré poskytujú a technológiami, ktoré používajú.

Google Maps

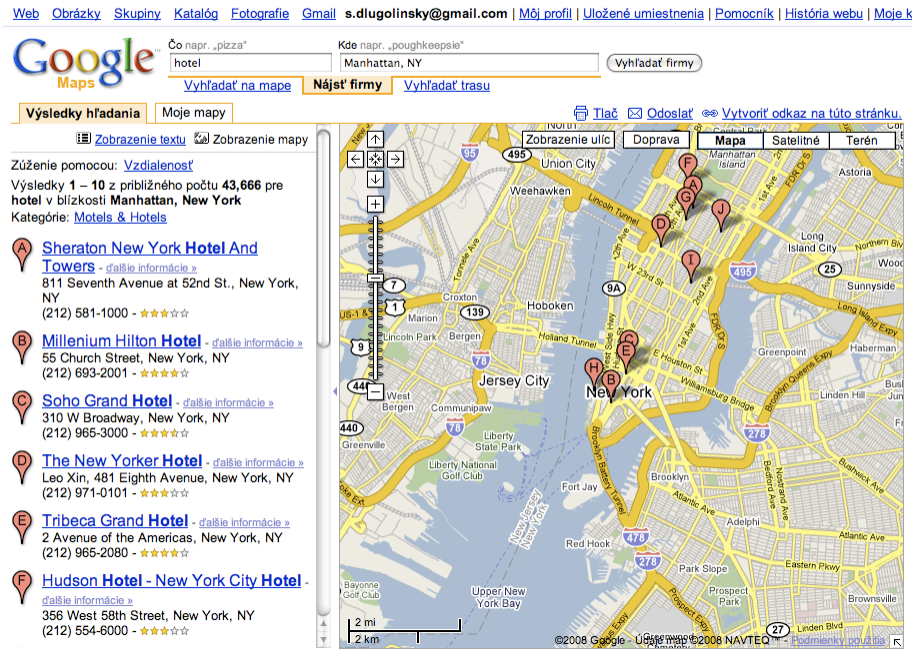
Google Maps[1] je projekt spoločnosti Google. Ide o bezplatnú internetovú službu určenú predovšetkým na vyhľadávanie geografických lokalít po celom svete. Okrem toho poskytuje možnosť hľadania firiem a vyhľadania trasy. Mapy sa dajú zobraziť v troch režimoch: klasická mapa, satelitný snímok a terén. Na prenos geografických údajov medzi aplikáciami bol vytvorený formát KML.

Vyhľadávanie

Vyhľadávanie prebieha zadaním dopytu do vyhľadávača. Výsledky vyhľadávania možno obmedziť na lokality, firmy alebo obsah definovaný používateľom. Na Obr. 1 je výsledok vyhľadávania dopytu „hotel“ medzi firmami v lokalite „Manhattan, NY“. Vyhľadávanie vrátilo 10 výsledkov z približného počtu 43 666. Ako vidno aj z obrázka, výsledky sa nezoraďujú podľa vzdialenosti. Nájdené výsledky je však možné zúžiť do okruhu 90, 45, 15, 5 míľ, 1 míle a 2500 stôp. Na mapách sa dá vyhľadávať iba v obsahu definovanom používateľmi. Na pridanie vlastného obsahu do máp je potrebné sa prihlásiť pomocou konta Google do

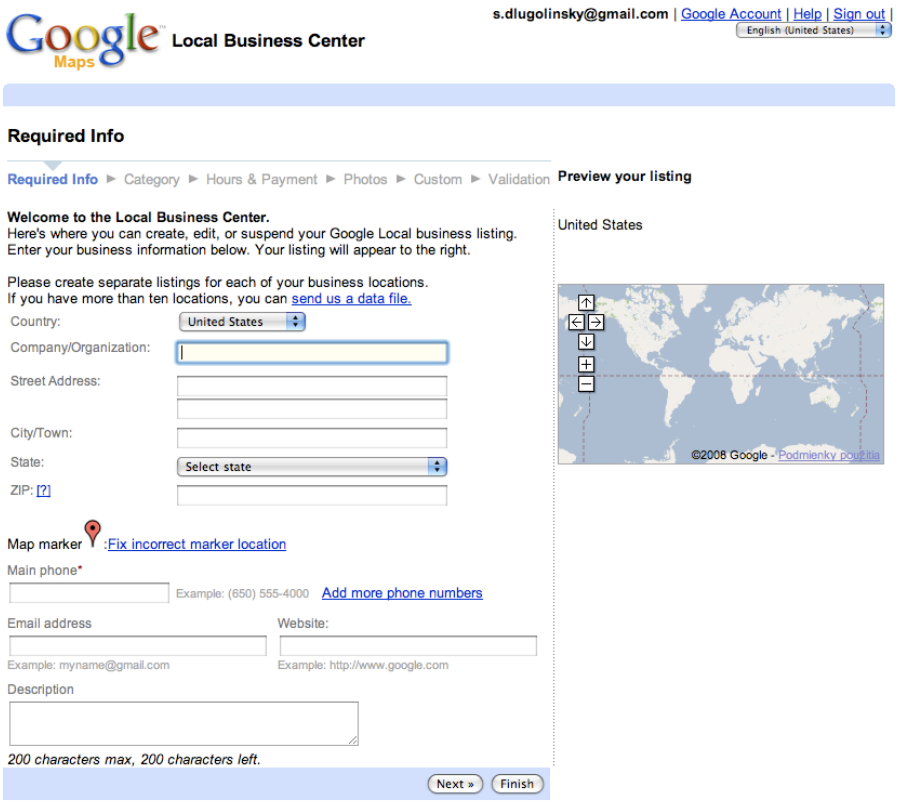
1 http://maps.google.com/

služby Local Business Center[2], kde je možné zadefinovať vlastné objekty (firmy poskytujúce nejaké služby) (Obr. 2). Pri pridávaní nového objektu na mape treba uviesť názov firmy, presnú adresu s poštovým smerovacím číslom, telefón a popis (max. 200 znakov). V ďalšom kroku sa vyberie kategória, pod ktorú bude objekt patriť. Ďalej je možné uviesť otváracie hodiny, spôsob platenia, pridať fotografie a videá z portálu YouTube.com. Posledným krokom je verifikácia, kde sa prostredníctvom automatického telefonického hovoru, alebo sms správy overí validačný kód. Zatiaľ nie je možné pridávať firmy so sídlom na Slovensku, pretože podporovaných je iba niekoľko krajín. Medzi susedov Slovenska, ktorí tu sú podporovaní, patrí Česká republika, Rakúsko a Poľsko.



Obr. 1 Google Maps

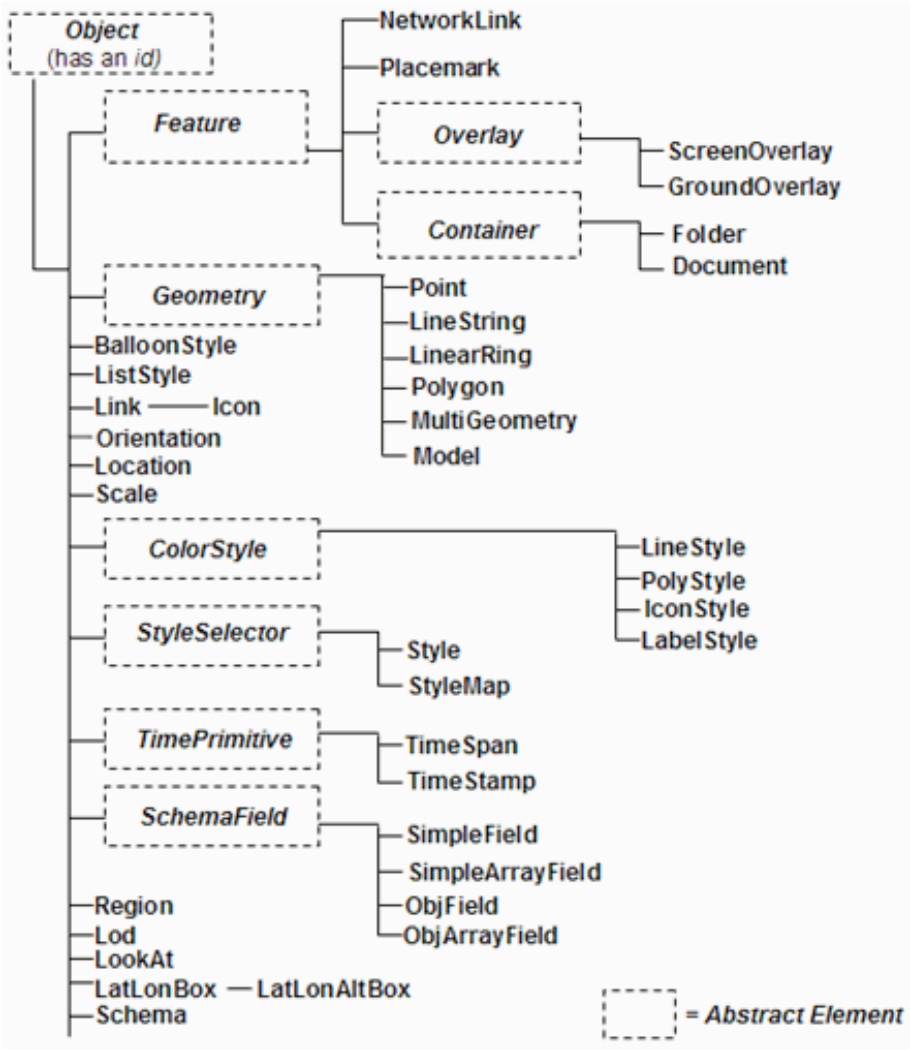
[2] http://www.google.com/local/add/lookup?hl=en-US&gl=US



Obr. 2 Pridávanie vlastných objektov do máp Google

Formát KML

Je to formát súborov používaných na uchovanie geografických údajov [6]. KML vychádza zo štandardu XML a v súčasnosti sa používa verzia KML 2.1. V beta testovaní je však už verzia 2.2. Súbory KML je možné otvoriť v prehliadači Google Earth, Google Maps a Google Maps for Mobile. Na Obr. 3 je znázornená štruktúra KML súboru verzie 2.1. Z hľadiska určovania geografickej polohy je zaujímavý hlavne element Placemark. Element Placemark obsahujúci element Point, určuje geografické miesto na mape. Okrem elementu Point môže element Placemark obsahovať ďalšie užitočné elementy ako Name, Address či AddressDetails. Rozdiel medzi Address a AddressDetails je v tom, že AddressDetails je štrukturovaná adresa formatovaná podľa medzinárodného štandardu xAL (eXtensible Address Language), kde Address je reťazec pozostávajúci z názvu ulice, mesta, okresu/štátu a/alebo smerového čísla.



Obr. 3 Štruktúra KML 2.1

Google Geocode

Geokódovanie (v anglickej literatúre označované geocoding alebo geocode) je automatický proces prideľovania zemepisných súradníc zadaným adresám [5]. Napríklad adrese „1600 Amphitheatre Parkway, Mountain View, CA“ bude pridelená zemepisná šírka 37.423021 a zemepisná dĺžka -122.083739. Na využívanie geocode služby poskytovanej spoločnosťou Google je potrebné mať Google účet a prihlásiť sa do Google Maps API služby. Ďalej si treba vygenerovať Google Maps API kľúč[3], ktorý sa bude viazať na konkrétnu stránku, kde sa bude Google Maps API používať. Limit na počet zobrazených stránok používajúcich Google Maps API nie je stanovený, ale pokiaľ sa očakáva viac ako 500 000 zobrazení denne, je odporúčané kontaktovať Google. Pre geocode službu je limit 15 000 požiadaviek denne. Okrem toho nesmie byť služba využívajúca Google Geocode spoplatnená koncovým používateľom. K službe Google Geocode možno pristupovať cez HTTP požiadavky alebo GClientGeocoder

[3] http://code.google.com/apis/maps/signup.html

objektom z Google Maps API. Treba však povedať, že Google Maps API je implementované v JavaScripte a je určené na vloženie do internetových stránok.

Údaje o geografických kódoch, používané v obsahu máp v službe Google Maps, sú poskytované prostredníctvom spoločností Navteq North America LLC (NAVTEQ), Tele Atlas North America, Inc. (TANA) a ďalšími nezávislými dodávateľmi. Na tieto údaje sa vzťahujú zmluvné podmienky ich používania. Používaním služby Google Maps a akýchkoľvek údajov, alebo informácií získaných pomocou tejto služby, sa vyslovuje súhlas s dodržiavaním zmluvných podmienok služieb Google a ďalších dodatočných podmienok[4]. V podmienkach sa okrem iného uvádza, že nie je možné používať službu Google Maps spôsobom, ktorý umožní používateľovi alebo inej osobe pristupovať k hromadnému preberaniu alebo získaniu veľkého množstva číselných súradníc zemepisnej dĺžky a šírky.

Geocoding cez HTTP

Goeocode požiadavka sa odošle na adresu *http://maps.google.com/maps/geo?* s nasledovnými parametrami v URI:

- q – adresa, ktorá sa bude geokódovať
- key – Google Maps API kľúč
- output – formát výstupu (xml, kml, csv, json)

Ukážka HTTP požiadavky na geokódovanie reťazca „jasovská 49“:

http://maps.google.com/maps/geo?q=jasovsk%C3%A1+49&output=csv&key=ABQIAAAACsXoLKcKH05sHY44-ubZdRQgDvrvmodzu1MMiwayagfuvjzJwxRlsuND0klcs0WMzcCxQL02filkVQ

Výsledok požiadavky bol:

```
200,8,48.095227,17.119740
```

čo znamenalo úspešné geokódovanie.

Geocoding cez Google Maps API

Geocoding sa v Google Maps API dá realizovať aj pomocou objektu GClientGeocoder. Na konvertovanie reťazca adresy do geografických koordinátov je určená metóda getLatLng() vracajúca koordináty v objekte GlatLng. Metóda má dva parametre, adresu na geokódovanie a callback funkciu, ktorá sa zavolá po prijatí výsledku odoslanej požiadavky. GClientGeocoder sa dá nastaviť tak, aby preferoval výsledky z určitej lokality. Urobí sa tak metódou setViewport(), ktorej sa ako parameter zadá objekt GlatLngBounds špecifikujúci lokalitu. Ďalej je možné pomocou metódy setBaseCountryCode() obmedziť výsledky z vybranej krajiny. Z objektu GclientGeocoder sa dajú extrahovať štrukturované adresy metódou getLocations(). Tá vracia JSON objekt obsahujúci

[4] http://www.google.com/terms_of_service.html

tieto informácie:

- Status
  - request – typ požiadavky (v tomto prípade vždy geocode)
  - code – návratový kód požiadavky informujúci o jej úspešnosti
- Placemark – výsledky geokódovanej adresy
  - address – čitateľne sformátovaná adresa
  - AddressDetails – adresa vo formáte štandardu xAL
    - accuracy – presnosť s akou sa adresa geokódovala
  - Point – súradnice miesta v priestore
    - coordinates – zem. dĺžka, zem. šírka a výška miesta

GClientGeocoder je na strane klienta vybavený vyrovnávacou pamäťou výsledkov geocodera. Pri opätovnom geokódovaní adries sa výsledky čítajú práve z tejto pamäte.

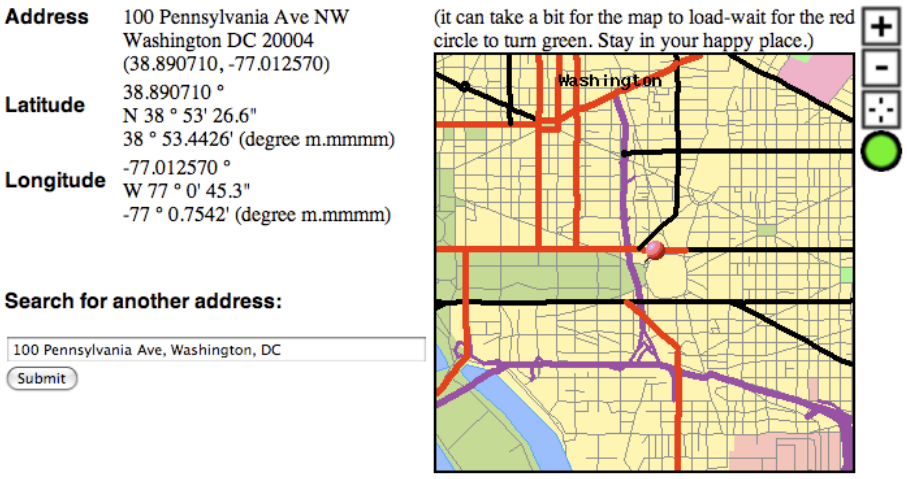
Zhodnotenie

Google Maps poskytuje komplexné služby vyhľadávania na vysokej úrovni. Výhodou je ich bezplatné využitie na nekomerčné účely, avšak s určitými obmedzeniami. Obmedzenia sa týkajú hlavne množstvom prenesených dát. Služba Google Geocode poskytuje veľmi presné zistenie zemepisných súradníc adries vrátane tých slovenských. Google Maps zatiaľ nemá vyriešené usporiadanie vyhľadaných výsledkov podľa vzdialenosti a samotné vyhľadávanie firiem na mapách je realizované v obsahu zadávanom používateľmi. Google teda automaticky nezisťuje lokalitu stránok.

Geocoder.us

Je to bezplatná geokódovacia služba[5] (pre nekomerčné účely), ktorá zistí zemepisnú šírku a zemepisnú dĺžku k zadanej adrese [4]. Geokódujú sa však iba adresy Spojených štátov amerických. Adresy musia obsahovať číslo ulice, názov ulice, mesto a štát alebo ZIP. Na Obr. 4 je ukážka geokódovania adresy „100 Pennsylvania Ave, Washington, DC“. Ku geokódovacej službe sa dá pristupovať pomocou rozhrania XML-RPC, SOAP, REST vracajúce RDF/XML dokument a REST vracajúce CSV súbor. Geocoder.us poskytuje aj parsovanie adries, ale nie ich štandardizáciu. Na hlavnej stránke sú k dispozícii aj zdrojové kódy programov komunikujúcich so službou Geocoder.us. Implementované sú varianty ASP/REST, PHP/SOAP, PHP/PEAR, PHP a C#. Ďalšou službou, ktorú Geocoder.us poskytuje je výpočet vzdialenosti medzi dvoma súradnicami alebo ZIP kódmi. Táto služba je bezplatne dostupná raz za 15 sekúnd. Platiaci klienti nemajú počet dopytov obmedzený.

[5] http://www.geocoder.us/



Obr. 4 Geokódovanie službou Geocoder.us

Zhodnotenie

Geocoder.us poskytuje služby na nekomerčné účely bezplatne. Nikde na stránke nie je presne uvedené, aké sú obmedzenia používania tejto služby. Prevádzkovatelia Geocoder.us si však vyhradzujú právo na reguláciu používania služby tak, aby bola vždy dostupná pre platiacich klientov. Platiaci klienti môžu využívať služby na komerčné účely. Čo sa týka obmedzenia geokódovania, tak to je ohraničené len na adresy zo Spojených štátov amerických.

Yahoo! Maps Web Services - Geocoding API

Spoločnosť Yahoo! taktiež poskytuje geokódovaciu službu, ku ktorej sa dá pristupovať pomocou Geocoding API rozhrania [14]. Geokódovanie prebieha vytvorením REST dopytu na adresu služby [15]:

http://local.yahooapis.com/MapsService/V1/geocode

Adresy Yahoo! služieb sa skladajú z niekoľkých častí. Prvá časť adresy je názov hostiteľa a pre službu geokódovania to je:

http://local.yahooapis.com

Nasleduje názov služby a jej verzia:

/MapsService/V1/

Za názvom a verziou služby sa zadáva názov metódy končiaci otáznikom:

geocode?

Ďalej sa uvádzajú parametre danej metódy, ktoré sú pre geokódovaciu službu uvedené v Tab. 1.

Parameter	Hodnota	Popis
appid	string (required)	ID applikácie pristupujúcej ku službe.
street	string	Názov ulice, číslo nie je povinné.
city	string	Názov mesta.
state	string	Názov štátu v USA alebo jeho kód.
zip	integer or <integer>-<integer>	Päť miestny zip kód, alebo päť miestny kód so štvormiestnym rozšírením. Ak je kód v rozpore s parametrom mesta (city) a štátu (state), na určenie polohy sa vezme iba parameter ZIP kódu (zip).
location	free text	Sem sa môže zadať jedna z nasledovných možností: • mesto, štát • mesto, štát, ZIP kód • ZIP kód • ulica, mesto, štát • ulica, mesto, štát, ZIP • ulica, ZIP Ak je nastavený tento parameter, bude mať vyššiu prioritu pri určovaní polohy ako ostatné parametre dopytu. Parametre city, state a zip budú ignorované.
output	string: xml (default), php	Formát výstupu.

Tab. 1 Parametre Yahoo! geocode REST dopytu

Príklad REST dopytu na geokódovaciu službu Yahoo! je nasledovný:

http://local.yahooapis.com/MapsService/V1/geocode?appid=YD-vdVay6Y_JXw.NWFSKRcPTg--&street=Jasovská+49&city=Bratislava

Pre horeuvedený dopyt bol vrátený výsledok

```
<?xml version="1.0"?>
<ResultSet xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="urn:yahoo:maps" xsi:schemaLocation="urn:yahoo:maps
http://api.local.yahoo.com/MapsService/V1/GeocodeResponse.xsd">
      <Result precision="address">
            <Latitude>48.095195</Latitude>
            <Longitude>17.119543</Longitude>
            <Address>49 Jasovska</Address>
            <City>851 07 Bratislava V</City>
            <State>Slovakia</State>
            <Zip></Zip>
            <Country>SK</Country>
      </Result>
</ResultSet>
<!-- ws06.search.re2.yahoo.com compressed/chunked Thu May  8
09:30:26 PDT 2008 -->
```

Pole	Popis
ResultSet	Obsahuje všetky vrátené výsledky
Result	Obsahuje každý jeden výsledok. Ak je adresa nejednoznačná, môže byť vrátených viac výsledkov. Atribúty výsledku sú: • precision: Presnosť adresy použitej pri geokódovaní v závislosti na jej extrahovaní z požiadavky. Možné hodnoty sú: o address o street o zip+4 o zip+2 o zip o city o state o country • warning: Ak nebola nájdená presná adresa, tu sa poznačí najbližšie nájdená.
Latitude	Zemepisná šírka
Longitude	Zemepisná dĺžka
Address	Adresa ulice výsledku, ak bolo možné určiť polohu.
City	Mesto, v ktorom sa výsledok nachádza
State	Štát, v ktorom sa výsledok nachádza
Zip	ZIP kód, ak je známy
Country	Krajina, v ktorej sa výsledok nachádza

Tab. 2 Návratové polia Yahoo! geocode REST dopytu

Zhodnotenie

Denný limit z jednej IP adresy na Yahoo! geocode službu je obmedzený na 5 000 dopytov za 24 hodín. Tento limit je možné prekročiť, ak pristupujeme k službe s dynamicky sa meniacou adresou. Ale takisto sa môže stať, že táto služba nebude dostupná, ak niekto vyčerpal limit a nám bola pridelená jeho IP adresa. Ako vidno na príklade vyššie, táto služba funguje aj so slovenskými adresami.

Nutch plugin GeoPosition

Ide o plugin do internetového vyhľadávacieho systému Nutch postavenom nad knižnicou Apache Lucene [7]. Plugin je schopný parsovať z dokumentov geografické meta tagy (geo.position, DC.coverage.spatial a ICBM). Ak sa z dokumentov žiadne geografické koordináty neextrahujú, môžu byť načítané zo súboru conf/geodata.txt. Príklad obsahu súboru geodata.txt:

```
http://www.berlin.de 52.1234 9.9876
http://www.germany.de/berlin 52.1234 9.9876
```

Na začiatku každého riadku je URL dokumentu a za URL je tabulátorom oddelená zemepisná šírka a dĺžka. Kladné čísla určujú severnú šírku a východnú dĺžku. Južná zemepisná šírka a západná zemepisná dĺžka je udávaná v záporných číslach.

Súradnicový systém

GeoPlugin vychádza z modelu Zeme ako gule s priemerom 6367 km, pričom maximálna chyba pri výpočtoch je autorom pluginu uvádzaná okolo 0.3 %. Taktiež predpokladá, že hladina mora je všade rovnaká. Pred samotným indexovaním sú geografické koordináty prepočítané do karteziánskeho súradnicového systému pozostávajúceho z osí x, y a z. Os x prechádza greenwichským poludníkom a rovníkom a os z severným pólom.

Ukladané dáta

Koordináty sú ukladané v polárnej sústave (severná zemepisná šírka a východná zemepisná dĺžka), ale neindexujú sa. Indexujú sa karteziánske koordináty, ale neukladajú sa a ak je k dispozícii aj výška nad hladinou mora (v metroch), uloží sa a indexuje.

Vyhľadávanie

Vyhľadávanie funguje na princípe položenia kocky okolo bodu v ktorom sa vyhľadáva. Ak sa na to pozrieme z dvojrozmerného pohľadu, tak sa pri vyhľadávaní okolo bodu berú do úvahy výsledky vo štvorci a nie v kruhu. To znamená, že maximálna odchýlka vo vzdialenosti je okolo 41.42 % a rozdiel v ploche pokrytia je okolo 27.32 %. Výhodou tohoto spôsobu je napriek vysokým odchýlkam rýchlosť. Odchýlky by mohol odstrániť tzv. distance ranking, ohodnotenie podľa vzdialenosti. Problém je však v zmene referenčného bodu. To znamená, že pre všetky body by sa muselo pri jeho zmene prepočítať ohodnotenie podľa vzdialenosti.

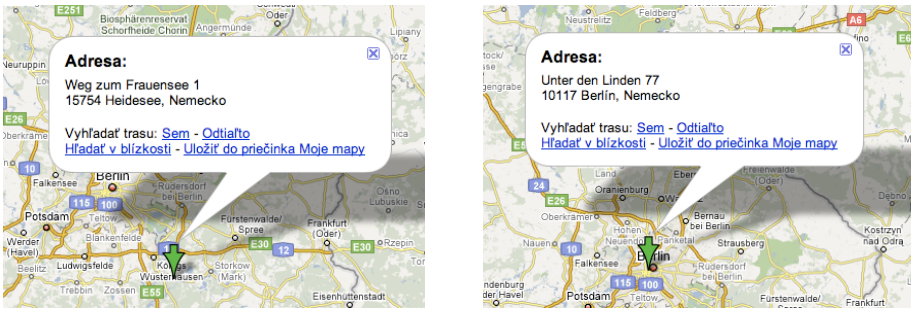
Inštalácia

Inštalácia prebieha nakopírovaním GeoPosition modulu do adresára sťahovača a servera (Apache Tomcat) a aktivovaním v konfiguračnom súbore nutch.conf. Kompletný návod na inštaláciu sa nachádza v inštalačnom archíve[6]. Posledná verzia modulu je verzia 0.5 z roku 2005.

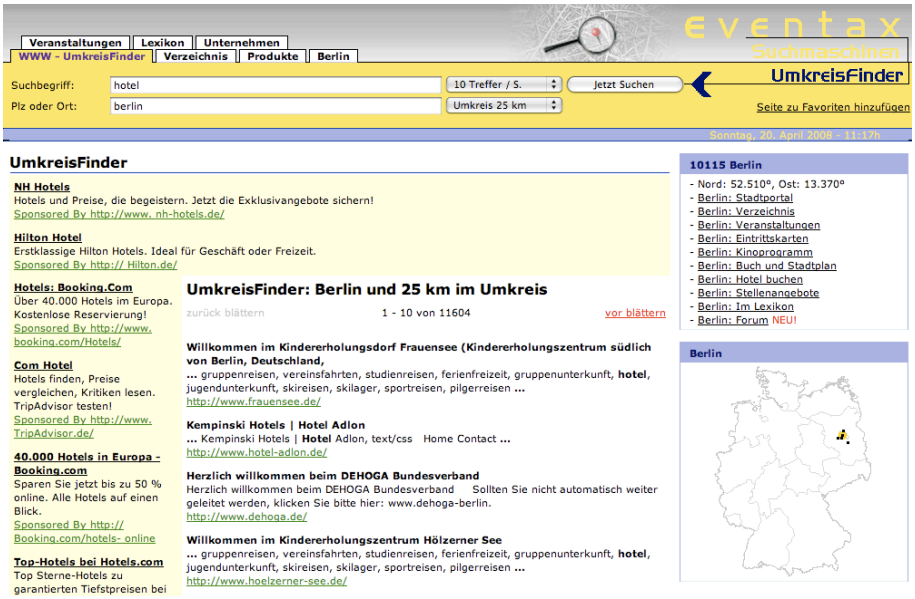
Použitie

GeoPosition plugin pre Nutch je použitý na nemeckom vyhľadávacom portáli UmkreisFinder.de[7] (Obr. 6). Dajú sa tu vyhľadávať dokumenty z nemeckých stránok. Výsledky vyhľadávania sú zobrazené aj na mape. Pri vyhľadávaní sa dá zvoliť referenčný bod (na obrázku Plz oder Ort). Potom sa hľadajú dokumenty v okolí tohoto bodu. Výsledky však nie sú usporiadané podľa vzdialenosti. Pre dopyt „hotel“ a referenčný bod „berlin“ vyhľadávač vrátil 11 604 výsledkov. Ako prvý bol vo výsledkoch uvedený hotel KiEZ Frauensee s adresou Weg zum Frauensee 1, 15754 Heidesee OT Gräbendorf. Ten je od Berlína ďalej ako druhý nájdený hotel s adresou Hotel Adlon Kempinski Berlin, Unter den Linden 77, 10117 Berlin, ktorý je priamo v Berlíne (Obr. 5).

[6] http://nutch.eventax.com/
[7] http://www.umkreisfinder.de/



Obr. 5 Porovnanie vzdialeností výsledkov v Google Maps



Obr. 6 Vyhľadávanie pomocou pluginu GeoPosition na UmkreisFinder.de

Zhodnotenie

Podľa autora GeoPosition modulu je potrebné skontrolovať existujúcu implementáciu, zrýchliť modul doimplementovaním vyrovnávacej pamäte pre dopyty a integrovanie s triedami FieldCache a HitCollector (balík org.apache.lucene.search). Okrem toho treba do modulu dorobiť ohodnocovanie dokumentov podľa vzdialenosti a parsovanie adries a zistenie zemepisných súradníc priamo z webovských dokumentov. Autor taktiež navrhuje, pokiaľ to prinesie dobrý výsledok, použitie služby whois na lokalizovanie dokumentu.

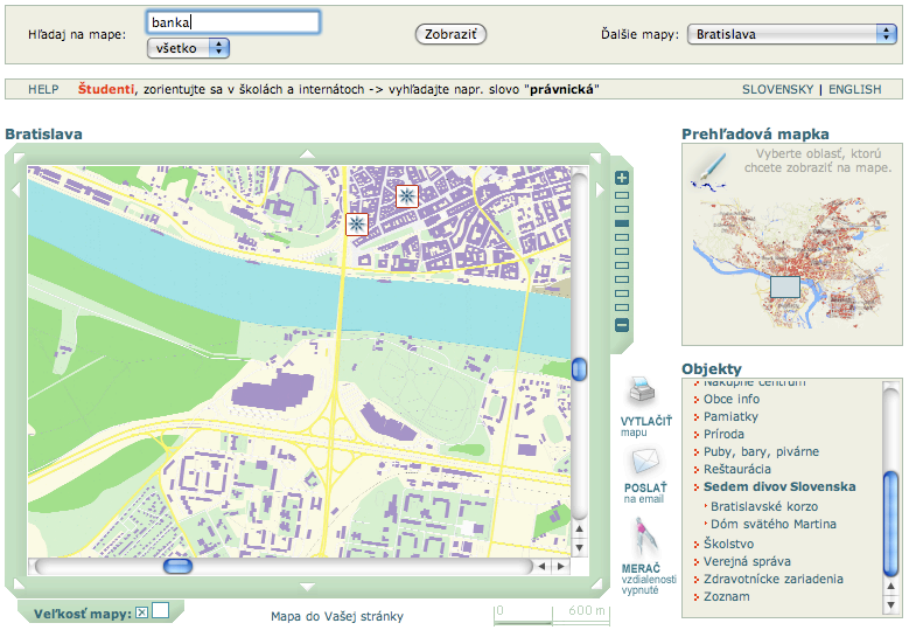
Kompas.sk

Na Slovensku poskytuje vyhľadávanie na mapách portál Kompas.sk[8] [19]. Je zameraný iba na Slovensko. Na mapách sa dajú vyhľadávať mestá, obce, konkrétne ulice a voliteľné objekty. V databáze vyhľadávania sa nachádza 138 miest, 2 891 obcí a okolo 2 000 štálov, lazov, názvov pohorí, riek a miestnych geografických názvov.

[8] http://mapy.zoznam.sk/

Vyhľadávanie

Výsledky vyhľadávania sú zoraďované podľa veľkosti sídla. Pri hľadaní ulice v konkrétnom meste sa zobrazí ulica v príslušnom meste, ďalej sú zoradené ulice s rovnakým názvom podľa veľkosti sídla. Vyhľadávanie objektov na mapách je možné pomocou názvu, napríklad Baumax, reštaurácia Rokoko, Aupark... Objekty, ktoré sa dajú na mapách vyhľadávať, boli do systému pridávané ručne, avšak tvorcovia portálu sľubujú prepojenie máp s katalógom Zoznam.sk. Doména objektov vyhľadávania na mape sa tak mala rozšíriť o firmy zaregistrované v katalógu. Zatiaľ sa dá vyhľadávať spomedzi 6 886 objektov ako sú zastávky MHD, kultúrne inštitúcie, pamiatky, reštaurácie a rôzne úrady a organizácie. Na Obr. 7 je ukážka vyhľadávania banky v Bratislave. Bolo nájdených 1 496 výsledkov. Vyhľadávač nebral do úvahy vyhľadávanie v zobrazenom výreze mapy a výsledky boli usporiadané podľa názvov nájdených objektov v abecednom poradí. Na Obr. 8 vidíme výsledky vyhľadávania banky v Bratislave. Medzi nimi sú aj také, ktoré sa nenachádzajú na miestach viditeľných na mape pri zadávaní dopytu.



Obr. 7 Vyhľadávanie na portáli Kompas.sk

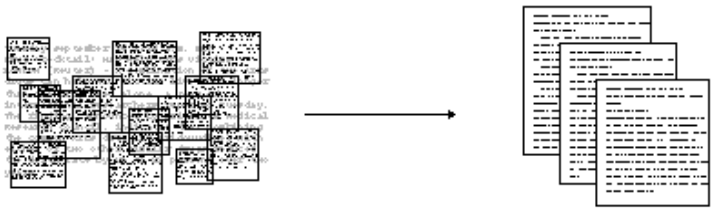
Obr. 8 Výsledky vyhľadávania na portáli Kompas.sk

Zhodnotenie

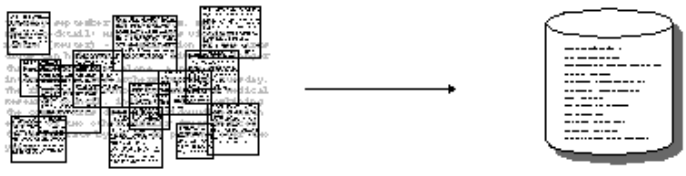
Vyhľadávanie portálu Kompas.sk je obmedzené len na zadefinované objekty vo svojej databáze. Vyhľadané objekty sa neusporiadavajú podľa vzdialenosti od referenčného bodu, ale abecedne podľa ich názvu.

Extrakcia informácií

Extrakcia informácií, v anglickej literatúre označovaná ako Information Extraction (IE), je proces analýzy textu s cieľom získať z neho štruktúrované dáta [2]. Takto získané dáta je ďalej možné podrobiť analýze, zobraziť používateľovi, alebo použiť na indexovanie v systémoch vyhľadávania informácií (Information Retrieval IR). Rozdiel medzi vyhľadávaním informácií a ich extrakciou je ten, že pri vyhľadávaní sa hľadá čo najrelevantnejší dokument z množiny dokumentov (Obr. 9), zatiaľ čo extrakcia informácií z dokumentov tej istej množiny získava iba požadované informácie (Obr. 10) [10].



Obr. 9 Vyhľadávanie informácií



Obr. 10 Extrakcia informácií

Extrakciou informácií sa zaoberajú konferencie MUCs (Message Understanding Conferences). Týchto konferencií bolo zatiaľ od roku 1987 dovedna sedem a definujú päť základných úloch extrakcie informácií:

1. **Named Entity recognition (NE)**
   Rozpoznávanie mien ľudí, názvov organizácií, adries, dátumov, finančných údajov a pod. Teda hľadanie pomenovaných entít v texte (názvov).

2. **Coreference resolution (CO)**
   Identifikovanie vzťahov medzi entitami. Rieši sa problém, kedy môže mať jeden objekt viac pomenovaní a tým pádom sa k nemu môže nájsť viac entít, napríklad zámien.

3. **Template Element construction (TE)**
   Hľadanie slov, ktoré bližšie charakterizujú pomenované objekty ako sú prídavné mená. Pri hľadaní sa využívajú entity CO, ktoré sú vo vzťahu s entitami NE.

4. **Template Relation construction (TR)**
   Hľadanie vzťahov medzi TE entitami. Napríklad hľadanie vzťahu medzi firmou a zamestnancom, alebo určenie príbuzenského vzťahu medzi dvoma ľuďmi a pod.

5. **Scenario Template production (ST)**
   Ide o spájanie výsledkov úloh TE a TR. Hľadajú sa tak fakty v súvislosti s entitami.

V Tab. 3 sú uvedené orientačné výsledky riešenia úloh extrakcie informácií, ktoré dosahujú súčasné systémy [12]. Údaje v tabuľke sú z meraní systému GATE pre anglický, španielský, japonský a čínsky jazyk. Presnosť úlohy ST nie je až taká nízka, ako sa na prvý pohľad zdá. Človek je schopný túto úlohu riešit s priemernou presnosťou okolo 80%.

Úloha	Presnosť
NE	97%
CO	60-70%
TE	80%
TR	75-80%
ST	60%

Tab. 3 Presnosť riešenia úloh extrakcie informácií súčasnými systémami

Prístupy extrakcie informácií

Na extrakciu informácií sa vo všeobecnosti používajú dva prístupy, využívanie znalostí a učenie. Znalostným inžinierstvom sa extrakcia rieši pomocou pravidiel. Nevýhoda tohoto prístupu je v dlhom vývoji systému a takisto sa v ňom ťažšie robia niektoré zmeny. Na druhej strane systémy postavené na tomto prístupe dosahujú lepší výkon v porovaní s učiacimi sa systémami. Tie používajú na učenie štatistické metódy a strojové učenie. Majú výhodu v tom, že sa v nich dajú

pomerne ľahko robiť úpravy, ale naopak, na ich učenie treba veľkú množinu trénovacích dát.

V súvislosti s automatickým zisťovaním geografickej polohy webovského dokumentu podľa jeho obsahu, ide o prvú úlohu extrakcie informácií (NE). V obsahu dokumentu treba hľadať entity, ktoré by sa dali využiť pri geokódovaní. Také entity môžu byť napríklad adresy, názvy ulíc, námestí, názvy miest, PSČ, smerové telefónne čísla alebo priamo geografické súradnice a súradnice GPS. Veľa firiem používa na svojich stránkach Google mapy na zobrazenie polohy svojich sídiel. Z týchto máp by sa taktiež dali extrahovať geografické súradnice a adresy. Pred samotným hľadaním entít však treba dokument najprv predspracovať (pre-processing). Text dokumentu, z ktorého sa informácie extrahujú, môže mať rôznu podobu. To v sebe skrýva veľa nástrah. Môže obsahovať formátovacie znaky, HTML kód, nemusia byť použité interpunkčné znamienka, kódovanie znakov môže byť rôzne, slová môžu obsahovať gramatické chyby, alebo môžu byť rozdelené a pod. Taktiež aj obsah textu alebo jazyk môže mať vplyv na extrakciu informácií. Všeobecne sa predspracovaním textu vykonávajú nasledovné úlohy:

- určenie typu dokumentu
- tokenizácia
- segmentácia slov (word segmentation)
- rozlíšenie významu homonymných slov (sense disambiguation)
- rozdeľovanie viet (sentence splitting)
- určovanie slovných druhov (POS tagging)

Po predspracovaní dokumentu prichádza na rad rozpoznávanie entít.

Slovníkový prístup

Najjednoduchší spôsob rozpoznávania entít (NE) je pomocou slovníku. Vyparsované tokeny z textu sa porovnávajú so vzormi v slovníku. Tento prístup je rýchly, ale na druhej strane závisí presnosť rozpoznania entít od obsahu slovníka. V slovníku môžu byť uložené napríklad názvy miest, obcí a ulíc. Pre predstavu, v Mestskej časti Bratislava – Staré Mesto je okolo 400 ulíc a námestí. Na intenetovej stránke Slovenskej pošty je dostupný súbor so všetkými smerovacími číslami ulíc na Slovensku a takisto súbor so všetkými názvami obcí a miest Slovenska[9]. V slovníku by mohli byť aj smerové telefónne čísla[10]. Podľa smerového čísla by sa dal určiť názov mesta. Parsovanie tokenov je možné riešiť regulárnymi výrazmi. Napríklad taký regulárny výraz na parsovanie GPS polohy by vyzeral nasledovne:

zemepisná dĺžka 0-180:

```
dlzka = "((0?0?[0-9]|0?[1-9][0-9])|1([0-7][0-9]|80))";
```

zemepisná šírka 0-90:

```
sirka = "(0?0?[0-9]|0?([1-8][0-9]|90))";
```

[9] http://www.posta.sk/storage/File/PSC/psc.zip
[10] http://telefonny.zoznam.sk/hladaj.fcgi?co=telzoznam&of=0&wht=smerove

minúty, sekundy 0-60:

```
minsec = "(0?[0-9]|([1-5][0-9]|60))";
```

stotiny 0.0-99.99:

```
stotiny = "(0?[0-9]|[1-9][0-9])[.,](0?[0-9]|[1-9][0-9])";
```

regulárny výraz na parsovanie GPS polohy:

```
"\\b" + sirka + "°" + minsec + "'" + stotiny + "\"[nN]([,;][\\s]*|[\\s]+)" + dlzka
+ "°" + minsec + "'" + stotiny + "\"[eE]\\b"
```

Tento regulárny výraz by vyparsoval z textu tieto GPS polohy:

```
48°16'55.40"N, 17°15'43.20"E
0°01'59.4"N,9°56'31.2"E
4°6'0.80"N;   179°01'21.80"E
```

Podobne sa dajú napísať regulárne výrazy pre PSČ, telefónne čísla a pod.

Shallow Parsing

Ide o techniku, ktorá sa snaží hľadať v texte štruktúrované informácie ako je napríklad adresa [11]. Štruktúra sa neparsuje ako celok. Parsuje sa iba jej časť. Pri adrese by tou časťou mohlo byť napríklad PSČ. Potom by sme predpokladali, že sa za PSČ nachádza názov mesta alebo obce a pred PSČ číslo ulice a jej názov. Takisto sa môžme pozrieť na názov ulice ako na štruktúru. V názvoch ulíc sa môžu vyskytovať slová: ulica, námestie, cesta, rad. Tiež môžeme predpokladať, že pred alebo za týmito slovami nasleduje zvyšná časť názvu adresy. Pre túto techniku existuje niekoľko modelov ako napríklad Hidden Markov Model, Generalized Winnow a Memory-based model. Zaujímavý je systém ALLiS, ktorý si podľa trénovacej množiny vytvorí strom pravidiel. Na základe stromu sa potom rozhoduje pri parsovaní štruktúr.

GATE

Jedným zo systémov, ktoré riešia úlohy extrakcie informácií je systém GATE (General Architecture for Text Engineering). Systém GATE je spolu so zdrojovými kódmi v jazyku Java voľne dostupný[11] pod LGPL licenciou. Tá dovoľuje jeho použitie aj v komerčných projektoch. Gate sa dá integrovať so širokou škálou ďalších systémov Tab. 4.

Oblasť	**Systémy**
Vyhľadávanie informácií	Lucene, Google, Yahoo!
Strojové učenie	Weka, SVMLight
Ontológia	Sesame, OWLIM
Parsovanie	RASP, Minipar, SUPPLE
Iné	UIMA, Wordnet, Snowbal, ...

Tab. 4 Podpora systému GATE

Systém GATE bol použitý v mnohých projektoch na extrakciu informácií. Išlo o projekty týkajúce sa bioinformatiky, medicíny alebo bezpečnosti. Tieto projekty patrili v medzinárodných súťažiach (MUC, ACE, Pascal) medzi najlepšie [9]. Štandardná distribúcia systému GATE obsahuje aj systém na extrakciu informácií ANNIE.

[11] http://sourceforge.net/projects/gate/

ANNIE

ANNIE (A Nearly-New IE system) je systém na extrakciu informácií [1]. Je súčasťou systému GATE, ale je možné ho použiť samostatne, alebo na vytvorenie nových aplikácií. Hlavnými súčasťami systému, ktoré riešia úlohy predspracovania textu sú:

tokenizér
Rozdeľuje text na jednoduché tokeny, ako sú čísla, špeciálne znaky, symboly a slová rôzneho druhu, napríklad s veľkými počiatočnými písmenami a pod.

rozdeľovač viet (sentence splitter)
Rozdeľuje text na vety. Tento modul je potrebný pre modul POS tagger.

POS tagger
Vychádza z Brill taggera. Spracovávaným slovám a symbolom priraďuje tagy, ktorými ich popisuje.

miestopisný zoznam (gazetteer)
Obsahuje zoznam miest, organizácií, dní v týždni a pod. Okrem týchto entít obsahuje aj rôzne skratky, napríklad obchodné skratky, či tituly. Zoznamy sú kompilované do konečných stavových automatov. Tie sa potom používajú na rozpoznávanie tokenov.

konečný stavový automat (finite state transducer)
Automat sa používa v moduloch rozdeľovača viet a miestopisného zoznamu.

orthomatcher
Hlavnou úlohou modulu je nájsť koreferenčné entity. Vie sledovať entity podľa rozpoznania ich vzájomného vzťahu. Ďalšou úlohou je tagovanie nerozpoznaných názvov, ktoré sa neskôr na základe vzťahov a tagov môžu rozpoznať.

coreference resolver (rieši úlohu CO).
Slúži na hľadanie vzťahov medzi entitami NE a na ich sledovanie.

Zhrnutie

V predchádzajúcich kapitolách sme si predstavili niekoľko existujúcich informačných systémov riešiacich geokódovanie a niektoré systémy aj s vyhľadávaním služieb na mapách. Báza dokumentov, respektíve objektov a firiem, v ktorej sa uskutočňovalo hľadanie, bola tvorená používateľmi. Výnimkou bol portál UmkreisFinder.de. Ten ponúkal vyhľadávanie aj v obsahu nezadanom používateľmi, teda medzi webovskými dokumentami. Umožňoval mu to systém Nutch vďaka Geoposition modulu. Modul však riešil geokódovanie len z časti. Vo webovských dokumentoch sa snažil nájsť meta tagy nesúce tri typy geograficej polohy. Súradnice teda nezisťoval z adries, ale ich priamo načítal z dokumentu. Ak nenašiel geografickú polohu v meta tagoch, tak ju skúšal nájsť vo svojom lokálnom zozname. Ďalej neriešil usporiadanie podľa vzdialenosti. V ďalšej práci by som chcel práve tento modul rozšíriť o extrakciu geografických údajov z obsahu dokumentov a vyskúšať geokódovať extrahované adresy. Čiže extrakcia a geokódovanie by boli mojim primárnym cieľom. Zameral by som sa na vyhľadávanie v ramci Bratislavy. Určite sa chcem pokúsiť aj o nájdenie riešenia pre usporiadavanie nájdených výsledkov podľa referenčného bodu, ale prioritu bude mať extrakcia a geokódovanie.

Zoznam použitej literatúry

[1] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications http://gate.ac.uk/sale/acl02/acl-main.pdf

[2] Cunningham, Hamish. 2004. Information Extraction, Automatic http://gate.ac.uk/sale/ell2/ie/main.pdf

[3] Fielding, Thomas Roy. 2000. Architectural Styles and the Design of Network-based Software Architectures http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

[4] Geocoder.us. 2006. Client Documentation http://geocoder.us/help/

[5] Google. 2008. Geocoding http://code.google.com/apis/maps/documentation/services.html#Geocoding

[6] Google. 2008. KML 2.1 Reference http://code.google.com/apis/kml/documentation/kml_tags_21.html

[7] Jaekle, Matthias. 2006. GeoPosition http://wiki.apache.org/nutch/GeoPosition

[8] Marshal, Dave. 1999. Remote Procedure Calls (RPC) http://www.cs.cf.ac.uk/Dave/C/node33.html

[9] Natural Language Processing Group, The University of Sheffield. 2008. Gate flyer http://gate.ac.uk/sale/gate-flyer/2007/gate-flyer-4-page.pdf

[10] Natural Language Processing Group, The University of Sheffield. 2008. Information Extraction http://gate.ac.uk/ie/index.html

[11] Stav, Adi. 2006. Shallow Parsing http://www.cs.tau.ac.il/~nachumd/NLP/shallow-parsing.pdf

[12] Tablan, Valentin. 2003. Information Extraction and GATE http://gate.ac.uk/sale/talks/manchester_11Dec03.pdf

[13] The Organization for the Advancement of Structured Information Standards (OASIS). 2002. Extensible Address Language (xAL) Standard Description Document for W3C DTD/Schema, Version 2.0 http://www.oasis-open.org/committees/ciq/download.html

[14] Yahoo! Inc. 2008. Yahoo! Maps Web Services - Geocoding API version 1 http://developer.yahoo.com/maps/rest/V1/geocode.html

[15] Yahoo! Inc. 2008. Creating a REST Request http://developer.yahoo.com/search/rest.html

[16] Berners-Lee, T., W3C/MIT, Fielding, R., Day Software, Masinter, L., Adobe Systems. 2005. RFC 3986 Uniform Resource Identifier (URI): Generic Syntax
http://www.ietf.org/rfc/rfc3986.txt

[17] RDF Core Working Group. 2008. Resource Description Framework (RDF)
http://www.w3.org/RDF/

[18] W3C XML Protocol Working Group. 2008. SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)
http://www.w3.org/TR/soap12-part1/

[19] Zoznam s.r.o.. 2008. Vyhľadávanie
http://www.zoznam.sk/help/?id=140