

Térinformatikai adatbázisok

Tanulmány az Adatbázisok haladóknak c. tárgyhöz

Sziroczák Richárd

2. évf. MSc Mérnök informatikus szak

2012/2013 tanév I. félév

Tartalomjegyzék

BEVEZETÉS.....	3
ELŐZMÉNYEK	3
TÉRKÉPÉSZETI BEVEZETŐ.....	3
MŰHOLDAS HELYZETMEGHATÁROZÁS BEVEZETŐ	6
TÉRINFORMATIKAI RENDSZEREK.....	6
SZEMANTIKUS HÁLÓ BEVEZETŐ.....	8
GOOGLE MAPS	10
GOOGLE DATA CENTER	13
GOOGLE BIGTABLE	14
ÖSSZEFOGLALÁS	15
IRODALOMJEGYZÉK	15

Bevezetés

A térinformatikai rendszerek a világ geográfikus adatait gyűjtik, feldolgozzák, tárolják és megfelelő megjelenítést valamint kimutatásokat készítenek. (Térképészet, meteorológia, stb.) A modern korral, az adatok digitalizálásával, a folyamatok automatizálásával és a különböző rendszerek integrálásával új dimenziók nyíltak meg a témában.

Mivel még nagyon széthúzó a világon az állapot a komoly térinformatikai rendszerekben, ezekkel csak említés szintjén foglalkozom. A mélyebben szemügyre vett téma ezen belül a helykereső és térképészeti szolgáltatások, amely egy dinamikusan fejlődő iparág, különösen a széles körben használt útvonaltervezés. Példaképpen a Google Maps-t mutatom be részletesen.

Előzmények

A téma főleg egy esettanulmány és a jelenkori helyzet bemutatása. A megértéséhez csak alapvető fogalmakkal kell tisztában lenni, amelyek egy felsőbb éves mérnökhallgatónak biztosan ismertek.

Térképészeti bevezető

A tanulmánynak nem feladata a térképészet tudományának megtanítása. Amit fontos megérteni a továbbiakban, az a probléma mélysége és a lehetséges módszerek tulajdonságai.

A térkép a világ egy vizuális reprezentációja, amelyen szimbólumokkal és kiemeléssel egyéb információt is kódolhatunk. Specifikusan egy földrajzi térkép a Föld felszínét ábrázolja, síkba vetítve.

Több különböző vetítési technika alakult ki a történelem folyamán és létezik napjainkban is. Minden technika alapja a Föld térbeli modellezése (gömb, valamilyen paraméterű forgás-ellipszoid, geoid), amelyről egy térből síkba vetítéssel, pontról pontra leképezés történik. A leképezés előtt és után transzformációkat is bevezethetünk, amelyekkel egyéb kívánatos tulajdonságokkal is felruházhatjuk a készülő térképet. (Például egybefüggő legyen, középpontjában egy adott hely legyen, stb.)

A Figure 1. ábrán láthatjuk az egyik legelterjedtebb és leginkább közismert vetítési módszert, a sztereografikus projekciót.

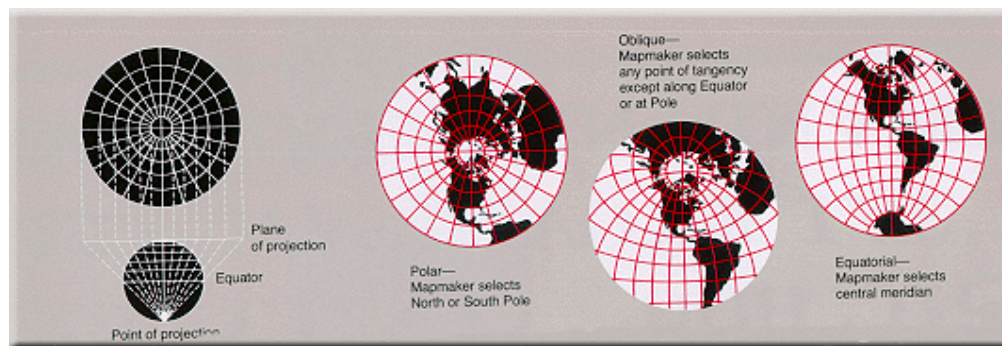


Figure 1: Sztereografikus projekció

Említésre méltó még a geoid (Figure 2. Ábra), amely a Föld valódi alakja, ellentétben egy idealizált forgás - ellipszoiddal. A geoid, Definíció szerint a nehézségi erő azon szintfelülete, ami a nyugalmi tengerszinttel esik egybe. Az eltérés oka a Föld-Hold keringési rendszer, az ebből következő ár-áply jelenség, a föld kérgében lévő eltérések, valamint a föld olvadt magjának keringése. A gyakorlati életben a használata nélkül maximum 200m hibát lehet ejteni, van ahol ez fontos, és van, ahol nem foglalkoznak vele.

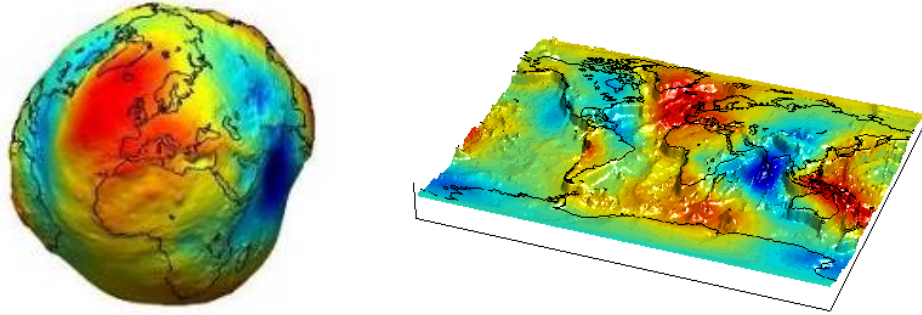


Figure 2: Geoid

A különböző vetítési technikák eredményének lehetnek hasznos tulajdonságai, amit elvárunk egy térképtől, ezek:

- Iránytartó (azimutális)
- Lokálisan alaktartó (konform)
- Területtartó
- Távolság tartó (egy vagy két ponthoz)
- Legrövidebb út tartó
- Az sem árt, ha összefüggő, konvex, vagy akár téglalap alakú az eredmény

Sajnos minden hasznos tulajdonság nem teljesülhet egyszerre. Például egy matematikai tétel kimondja, hogy nem lehet egyszerre konform és területtartó egy térkép. Láthatjuk például az Figure 1. ábrán, hogy a keletkező térkép nem iránytartó. (Ha az ember elindul dél felé egy hosszúsági kör mentén egyenesen megy a valóságban, egy ilyen térképen viszont nem egyenesen.)

A feladat vagy az igény dönti el, hogy milyen tulajdonságú térkép a leghasznosabb minden esetben. Legtöbbször elvárjuk, hogy területtartó legyen egy térkép (ami a térképen nagyobb részt foglal el, az nagyobb a valóságban is), néhány klasszikus vetítést lehet látni a Figure 3. ábrán.

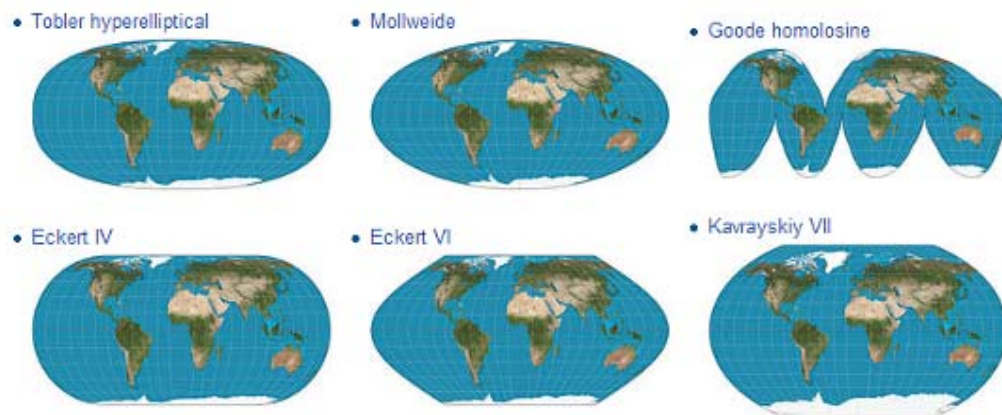


Figure 3: Klasszikus pszeudo henger vetítések

Lehet készíteni egy vagy két ponttól távolságtartó térképeket is. Az ilyen térképek leginkább Ázsiában elterjedtek. (Európában és Amerikában a térképek szélén lévő tengerek disztorciója nem számít sokat) Egy példát láthatunk a Figure 4. ábrán, amely a két jelölt piros ponthoz képest távolság tartó.



Figure 4: Két ponttól távolságtartó vetület

Még furcsábbnak látszik az az egyszerű kéréssel készített térkép, hogy legyen két pont között a legrövidebb út az egyenes. A Figure 5. ábrán láthatjuk a Gnomonic vetületet, amely az egyetlen elterjedt legrövidebb út tartó vetítési forma. Légi közlekedésben és hadászatban használják.

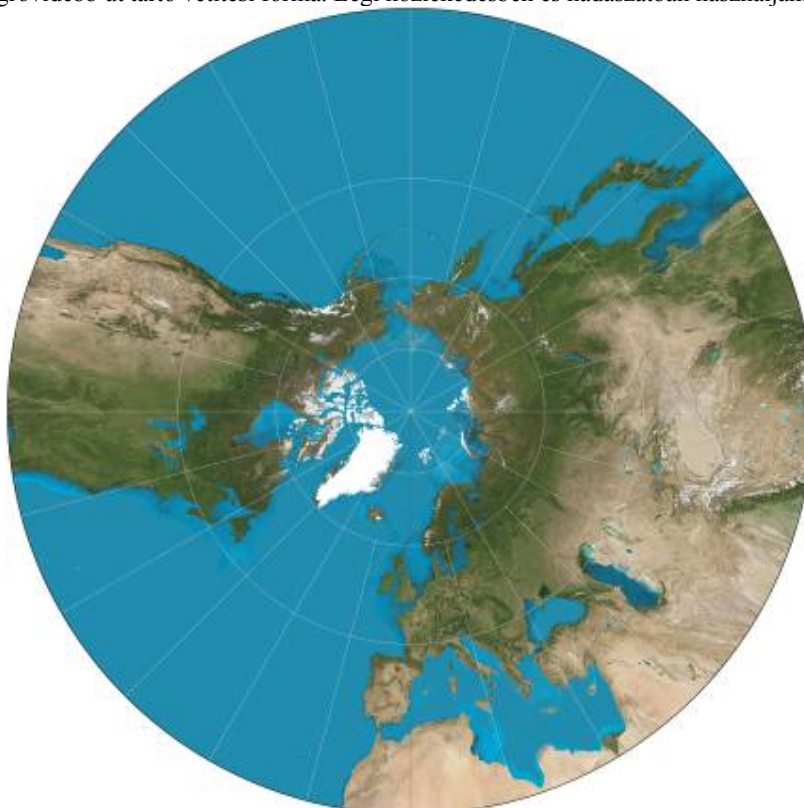


Figure 5: Legrövidebb út tartó vetítés

A bevezető tanulsága az, hogy ha egy térképet látunk, mindig hozzá kell érteni a térkép elkészítésének módját. Például ha egy körzővel a jelenlegi helyzetünket választva középpontnak húzunk egy kört egy térképen, az így kijelölt pontok elég valószínű, hogy nem alkotnak egy kört a valóságban.

Ugyancsak érdemes megemlíteni, hogy a legtöbb térkép, ami a mai napig használatban van, még a múlt századi technikákkal lett elkészítve (klasszikus helyszíni földmérés), a mai digitális korhoz képest nagyon rossz precizitással. Mivel a földmérés nagyon költséges folyamat, nem minden esetben beszélhetünk precíz modern térképekről.

Van úgy, hogy a térkép leolvasásából fakadó hiba elhanyagolható, de van, hogy olyan nagyságrendű, hogy nem. Például látni fogjuk, hogy egy navigáló szoftver esetén 50-100m hiba már nem elfogadható.

Műholdas helyzetmeghatározás bevezető

Az alábbiak abból a szempontból mutatják be a technológiát, hogy milyen szolgáltatást nyújthat egy informatikai rendszernek, amely ráépül.

A globális műholdas helyzetmeghatározó rendszer (GNSS - Global Navigation Satellite System) sokszor keveredik a hétköznapi használatban a GPS-sel (Global Positioning System). Az előbbi a technológia, az utóbbi egy konkrét rendszerszintű megvalósítása. Az Amerikai Egyesült Államok GPS rendszere az első ilyen rendszer a történelemben és 30 műholddal a mai napig szolgálja a NATO szövetséget valamint a fejlett országokban a magánszektor. Jelen pillanatban az egyetlen másik üzemelő globális lefedettségű rendszer az orosz GALONASS. Fejlesztés alatt áll az Európai Unió részéről a Galileo (2014) valamint Kínában a COMPASS (2020).

A rendszer lényege, hogy a Föld-Hold tömegközéppont körül meghatározott pályán kering több műhold, amely jeleket sugároz. 24-30 műholdból áll egy ilyen rendszer, amelyből minden pillanatban a Föld bármely pontján észlelhető 7-9 darab. Ebből 4-5 öt szoktak kiválasztani, azokat, amelyek jelei a legkevesebb atmoszférán haladnak át. A jel maga a műhold azonosítója egy időbélyeggel. Egy átlagosan precíz rendszerórával a Földön található vevő - készülék ki tudja számolni a távolságát az egyes műholdaktól (mennyi idő telt el a leadott időbélyeg és a vétel között ismert, jel terjedési sebessége ismert). Ez után ismerve a műhold pozícióját (ami publikus) valamint a Földön a domborzati viszonyokat mértani helyek metszésének módszerével meg lehet határozni három műholdból a földi vevő készülék pozícióját.

A gyakorlatban az algoritmus bonyolultabb, mert a zaj miatt pontatlanná válik a folyamat. Mégis, ha több műholdról is van adatunk és okos szűrést és adatfeldolgozást alkalmazunk, jó pontosság érhető el. Napjainkban a hardware követelménye, szoftveres számításigénye minimális egy ilyen földi vevőkészüléknek, szinte minden modern telefonban megtalálható.

Érdekesség, hogy az ultra precíz atomórával felszerelt műholdak konfigurálásánál konkrétan bele kellett számolni az idő dilatációt [1]. (A speciális relativitáselmélet egyik következménye a gyakorlatban is megjelent.)

Napjainkban elég elterjedt a földi vevőkről pillanatnyi adatot gyűjteni, feldolgozni és a tanulsággal javítani minden résztvevő további számítását. Ez általában néhány fix referencia pont, de van már működő példa a felhasználók között elosztottan is. Így az időjárési és egyéb zavaró tényezőket az időben folyamatosan minimalizálni lehet.

A GPS a napjainkban megfelelő segéd technológiákkal [2] elérheti a 0.4m –es pontosságot. (Pontosság alatt az átlagos sugárirányú szórást értendő.) Ez a szolgáltatás azonban csak a hadászati célokra használható és a hozzáférése titkosított. A magánszektori felhasználásra egy pontatlanabb jelet sugároznak a műholdak, amelyből maximálisan 5m -es pontosság érhető el. (Vagyis direkt zajt kevernek a jelre.)

Mindent összevetve ez a pontosság elegendő arra, hogy navigáló szolgáltatást építhessenek rá. A piacon kaphatók is kijelzős készülékek 100-200 USD árban, valamint alkalmazások okostelefonra 20 USD körül. A piac igen nagy és napjainkban is dinamikusan fejlődő.

Térinformatikai rendszerek

A térinformatikai rendszerek (Geographic Information System – GIS) [6] gyűjtik, tárolják, feldolgozzák, analizálják a földrajzi adatokat, amelyekből kimutatásokat és megjelenítéseket készítenek.

Egy térinformatikai rendszer létrehozása klasszikusan föntről lefelé történik. Egy megfelelő térkép elkészül (vagy vásárolnak), majd rendszeres adatgyűjtések keretében referencia pontokon megméri a világot. (Tehát a térkép pontjait keresik föl a valóságban és ott mérik.) A rendszer digitalizálása nem túl bonyolult feladat, maga az adatgyűjtés a nehéz és költséges folyamat, amelyet nem könnyű automatizálni.

Jelenleg csak igen nagy szervezetek és hivatalok működtetnek térinformatikai rendszereket, a saját feladatuk segítésére. Például a földhivatal térinformatikai rendszere nyilván tartja, mikor, hogyan volt a termőföld felparcellázva, az egyes területeken ki mit termesztett, a rendszeres talajminőség méréseknél milyen adatokat kaptak, stb. Ha az aktuális belvízi helyzetet szeretnénk megtudni, akkor már a katasztrófa és árvízvédelmi hivatalba kell menni, és az ottani térinformatikai rendszerből tudhatjuk meg az adatokat.

Feltett cél a jövőben, hogy komplex rendszereket hozzanak létre, ahol több, különböző témájú adat is kapcsolódik egy közös geográfiai reprezentációhoz. Hogy egy rendszerből megállapítható legyen, hogy egy adott földterületnek ki a tulajdonosa, milyen közművek mennek a föld alatt, milyen minőségű a föld, stb. Egy ilyen komplex (réteges) térképet láthatunk a Figure 6. ábrán.

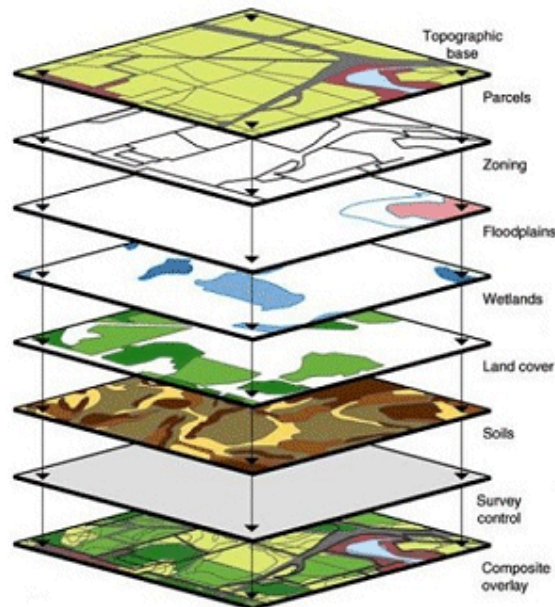


Figure 6: Komplex térkép

Ez viszont csak a jövő víziója [7], jelenleg nem túl kedvező a helyzet egy ilyen átfogó middleware gondolatbeli fejlesztésénél. A különböző térinformatikai rendszerek különböző típusú térképeket használnak, ezen különböznek a referencia pontok, más az adatgyűjtés frekvenciája, stb. Minden térinformatikai rendszer a kialakulásakor a saját feladatának megfelelően választotta ki ezeket a paramétereket, sokszor századokkal ezelőtt. Például azért nem lehet könnyű átállni más módszerre a meteorológiában, mert a századok alatt gyűjtött adatok hatalmas értéket képviselnek.

Tehát ha két különböző módszerekkel dolgozó térinformatikai rendszerből szeretnénk adatot összesíteni a következő igen nehéz problémákba futunk bele:

- Mi felel meg az egyik pontnak a másik térképen? Illetve a nehezebb változat, mi felel meg az egyik területnek a másikon?

Különbözhet a térképkészítés módszere. Különbözhet a referencia területek szervezése is. (Például az egyikben parcellák, a másikon földrajzi háló.) Egy ilyen helyzetet feloldani csak

bizonyos pontatlanságokkal lehet. A kemény probléma akkor keletkezik, ha a precizitás követelmény.

- Mit teszünk a hiányos adatokkal? Például az egyik rendszerben naponta, a másikban csak hetente mértek adatot.

Lehet esetleg átlagolni vagy időszorral becsülni. Ám az ilyen módszerek ismét pontatlanságot visznek a rendszerbe.

- Mit teszünk az ellentmondó adatokkal? Például ha az egyik rendszerben ennyi csapadékot mértek, a másikban annyit.

Nem is biztos, hogy a mérésben hiba volt, lehet, hogy az előbb említett pontok miatt keletkezik ellentmondás. (Az egyik helyen heti átlag egy öt négyzetkilométeres területre, a másikon napi átlag egy kisebb területre.)

Sajnos tehát el kell mondani, hogy a jelen kor térinformatikai rendszerei nagyon feladat-specifikusak, speciális módszereket használnak, és sokszor gúzsba vannak kötve a bürokráciával. Informatikai szempontból inkább az integrálásuk az izgalmas és nehéz feladat. Ebbe az irányba egyébként nagy költségvetésű projektek tartanak, hogy legyen egy átfogó middleware a különböző rendszerek között. (Sajnos itt sincs teljes összhang. Magyarországon ebben a pillanatban négy ilyen projekt is fut, ebből egy NATO, egy EU forrásból.)

Konkrétan adatbázisok szempontjából azt érdemes megemlíteni, hogy ezek a rendszerek nagyon precízen tervezettek. Az adatbázisok mögötte is a „minden lap előre leoszva” elv szerint készültek és működnek, amíg egy új rendszert nem terveznek a helyére.

Ezért is kormányozzunk egy közeli témára, ahol nagyon látványos eredmények születtek, és a mögötte lévő technikából is lehet sok újat tanulni.

Szemantikus háló bevezető

A Szemantikus háló (Semantic web, web of data) [8] egy koncepció, illetve módszertan az információ automatikus feldolgozásáról. A World Wide Web Consortium (W3C) szabványosította az elképzelést és az ajánlásait minden komoly szervezet követi. A szakmai körökben egyetértés van arról, hogy ez az Internet jövője, az élvadások már implementálták és így a szélesebb körben is (a trendkövetéssel) fokozatosan terjedni fog. (Szakzsargonban Web 3.0-nak is hívják ezt a víziót.)

Az információ automatikus feldolgozása, szövegértés, gépi látás nagyon nehéz feladat és nyitott kérdés a mai napig. Egy informatikai rendszernek strukturálni kell az adatokat, hogy megérthesse. Ezt manapság emberi erőforrásokkal lehet elvégezni, amely nagyon költséges. Nem is egyértelmű, mert az informatikai szakember nem feltétlenül érti az információt és az adott terület szakértője nem biztos, hogy megérti a strukturálás technikáját.

Ezen felül a valós életben sem várható el, hogy minden írást egy metanyelven végezzünk el vagy mellékeljünk egy strukturált adatbázist a mondanivalónk mellé. Pedig az emberiség összes írása rengeteg információt rejt, ha hozzáférhető lenne automatikus feldolgozással, nagy eredményeket lehetne elérni.

A szemantikus háló az abszolút minimumot célozza meg, amit ha egy ember követ, egy automatizált rendszer még megérti valamilyen szinten. Tehát ezek technikák, amelyet ha implementál a fejlesztő, használható formájú lesz a weblapja, dokumentuma. (Megjegyzendő, hogy az átlag felhasználó ebből annyival szembesül, hogy esetleg egy adott listából ki kell választania az irománya témáját. Ez azért elviselhető teher.) A Figure 7: Szemantikus háló. ábrán látható egy szemantikus háló réteges felépítése, tekintsük ezt alulról felfelé.

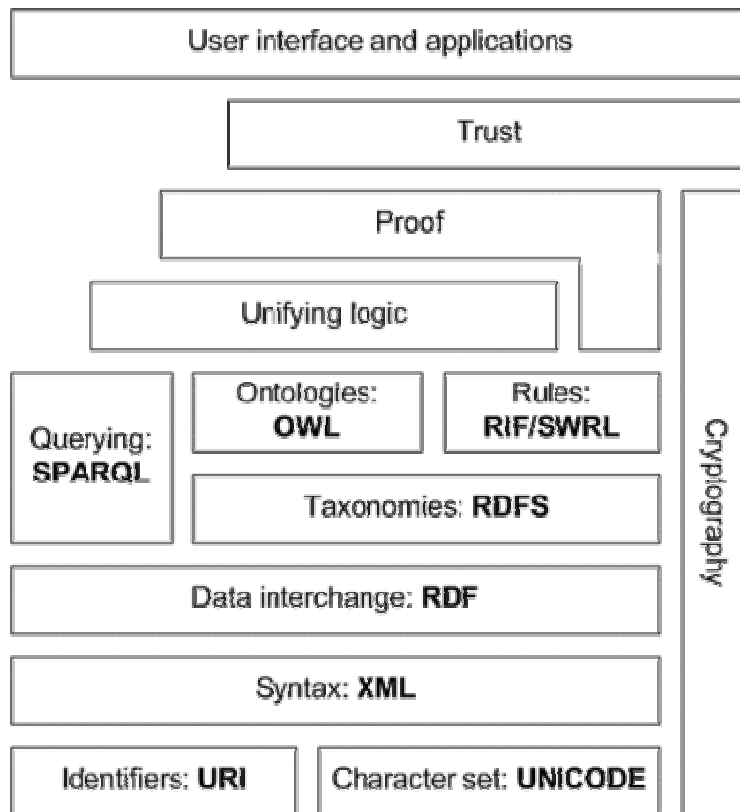


Figure 7: Szemantikus háló

Unicode-ot használjunk, teljes URI-kat és a dokumentumot XML szintaxissal kódoljuk. Egyértelmű, teljesíthető, amúgy is így csinálják a profik, ne is szaporítsuk a szót.

RDF (Resource Description Framework) [9] az adat közlés egy standard modellje. Fő erőssége, hogy az adatot úgy cserélhet két rendszer, hogy a sémáik különböznek. Illetve nem kell minden fogyasztót megváltoztatni, ha egy forrás sémája változik. Ezt úgy éri el, hogy az adathoz melléleli a sémát, amelyben jelzi az adat jelentését.

Ez még mindig a fejlesztő feladata. Például ha amúgy is egy adatbázisból dinamikusan generlja egy oldal tartalmát (például jelenlegi időjárás statisztikái), akkor már publikussá teheti az RDF sémát is, ami alapján generálta az oldalt. Így egy automatizált programnak nem a szabad szövegből kell információt leszűrnie, hanem direkte a forrásból, az adatbázisból meríthet.

A magasabb szint már nem feltétlenül a dokumentum fejlesztőjének a feladata, hanem az automatizált dokumentum olvasó rendszer fejlesztőjéé. Ehhez csak néhány információ morzsa szükséges, ami megjelöli a dokumentum témáját például, de ez sem feltétlenül szükséges. (Tanuló módszerekkel fel is lehet fedezni ezeket.) A témához az olvasó rendszer betölt egy megfelelő ontológiát és taxonómiát. Ez lehet saját maga által épített, vagy harmadik féltől vásárolt is. (A piacon profi cégek foglalkoznak ontológiák és taxonómiák gyártásával. Jelenleg kiélezett verseny folyik a nagyobb piaci részesedéssel.) Ezek segítségével a rendszer képes szabályszerűségeket felfedezni egy szabad szövegben. (Ezek leíró nyelve a RIF, és egy hozzáférési módja a SPARQL.)

Például. A szövegben az van, hogy „verjünk fel két tojás habját”. A szöveggörnyezetből meg lehet tudni (MI módszerek vagy akár csak szó, kifejezés gyakoriság alapján) vagy akár meg is adhatják, hogy itt egy receptről van szó. Témák: Vendéglátó ipar, főzés, tervrajz, stb. A rendszer betölti a főzés típusú ontológiát, amiből tudja, hogy egy tojás áll fehérjéből és sárgájából. A főzés típusú taxonómiából megtudhatja, hogy a tojás hab alatt valójában a tojás fehérje habbá vert formáját értjük. Így képes a rendszer egy strukturált formára hozni az instrukciókat, megértheti, mit írnak egy dokumentumban.

A következő magas szintű modellben már szabályokat lehet megtanulni a világról. Általános módszer, hogy ezekhez egy bizonyosságot rendelünk. Azok a szabályok, amelyek sok forrásból megerősítődnék bizonyosabbá válnak. A legbizonyosabb szabályokra lehet építeni, ezek a problémamegoldás módszereivé válhatnak.

Ez a szint nem szigorúan specifikált, a különböző feladatok különböző rendszereket követelnek meg. (Illetve ez még kutatási terület, nem alakultak ki egyértelmű szabványok.) Egy példát mutatok be a következőkben.

Kimondható tehát, hogy egy nagyon erős eszközt lehet felépíteni ezzel a módszerrel. Mivel az Internet és a digitalizált dokumentumok manapság bőven lefedik az összes emberi szellemi teljesítményt érdemes ezt a témát kutatni, hogy lehetne ezt a nagy értékű adathalmazt feldolgozni hatékonyan.

A továbbiakban egy olyan példát láthatunk, ahol ezzel a módszerrel próbálták meg feldolgozni a világot, majd ezt valós helyekhez is rendelték. Így egy újfajta megközelítést mutatva a térinformatikai rendszereknek.

Google Maps

Napjainkban a Google Maps [4] a fejlett világ minden országában teljes lefedettséggel működik a nap huszonnég órájában. Az évek alatt letisztult és hibáit kijavították. A maga nevében mindenképp piacvezető szerepet tölt be. Konkurencsei csak gyengébben tudják utánozni, de komoly költségvetéssel igyekeznek ezen.

Ugyan díjmentes a szolgáltatás, viszont a nagyobb üzleti tervben megéri fejleszteni, üzemeltetni. Ennek az okai:

- Remekül kiegészíti, emeli a keresőmotor értékét.
- Összefonódás az on-line navigáló rendszerekkel, valamint egyéb szolgáltatásokkal (Street view, tömegközlekedési menetrend). A jó logikai térkép nagy érték, eladható. Vannak olyan GPS alapú navigáló rendszerek, akik licenszelik és ráépítenek a Google Maps szolgáltatásaira.
- Ez is egy eszköz, hogy nagyobb szettet hasítsanak ki a modern személyi számítógépek piacáról. (Tablet, okostelefon, stb.)

A térkép felépítése eltér a hagyományos megközelítéstől. Itt az úgynevezett logikai térképet készítik el, és a megjelenítés csak egy eszköz, amire illesztik, majd „rágyúrák” a logikai térkép adatait.

Ez a rágyúrás (massage the data) egy sűrűn előforduló kifejezés az angol nyelvű forrásban [4] és leginkább azt az ad hoc, emberi intelligenciával végzett folyamatot takarja, amivel az egyes esetekben asszociálják az adatokat egy valós objektumhoz.

A logikai térkép objektum alapú alapigazságokból (Google Ground Truth) épül fel. Ezek az alapigazságok egy közös nagy szemantikus hálón helyezkednek el [8]. A Figure 7. ábrán látható a szemantikus háló elvi felépítése a W3C-től. Itt ezek az alapigazságok az egyesített logika szintjén helyezkednek el.

Például tudjuk, hogy van a Falu Széli Pékség nevű üzlet. Erről tudni lehet sok minden egyebet is, például a honlapján a Google keresőmotor elolvassa a metaadatokat. Feldolgozza az egész honlapot nyers szöveg alapján. Felismeri, hogy ez valami üzemi vagy vendéglátóipari témájú szervezet így megvizsgálja a szöveg jelentését abban az ontológiában, stb.

Felismerhet egy címet a szöveg között. Vagy kinézheti a cégnyilvántartásból is. (Bizonyíték tartozik mindenhez, azt az igazságot hiszi el, amire a legtöbb bizonyíték van.) Általában a cég a saját üzleti érdekeit (reklám, megtalálható legyen) szem előtt tartva segíti ezt a folyamatot.

Így máris asszociáció keletkezhet egy cím és egy fogalom között. Fontos látni, hogy amikor a felhasználó keres a rendszerben, ebben a hálóban keres.

Ilyen alapigazságként kezelik az utakat, útkereszteződéseket, épületeket, üzleteket, parkokat, látványosságokat, mindent. Köztük létrejöhet asszociáció, például ez az épület az adott út mellett van, az a bizonyos cég ebben az épületben van, stb. Sok adatot a hivatalos forrásokból merítettek, például a földhivatalból, cégnyilvántartásból [3]. Ezeknek az alapigazságoknak és a kapcsolatrendszerüknek az összessége a logikai térkép.

Hogyan lesz a logikai térképből egy valódi térkép?

Fordítva. Készítenek egy térképet és arra írják fel a fogalmakat [4].

A valódi térkép (az, amit a webes felületen láthatunk) műholdfelvételekből jön létre. Egy négyzetkilométeres ablakkal dolgoznak, 200m csúsztatással. Ezekből mindig az aktuális nagyítási szintnek megfelelő, pár felvételtől egyesített térkép látszik. (A nagyon messzi nagyítási szinten egy rajzolt térkép található. Megjegyzendő, az egyéb rajzolt térképek is a műholdfelvételeken alapulnak.) Ennek a térképnek az egyenetlenségeit javítani tudták az ismert domborzati viszonyokkal, de ennek nem volt túl nagy jelentősége. A kicsi, ablakozással hibajavított cellákon nagyon kevés a disztorció.

A valódi térképen először felismerik az utakat. Ezt egy jó minőségű felvételen meg tudják oldani a modern ipari képfeldolgozás módszereivel. A valóságban valamilyen tanuló algoritmus is jelen van ebben a folyamatban, ám komoly üzleti titokként kezelik. A lényeg ebben a lépésben, hogy az utak nagyját megtalálják a valós térképen. Hibák még bőven vannak ebben, mint az a Figure 8: Automatikus út felismerés. ábrán is látható a bekarikázott részeknél.



Figure 8: Automatikus út felismerés

A következő lépésnél történik a logikai térkép hozzárendelése.

A Google alapvető hozzáállása, hogy a több adatból jobb adatot lehet készíteni. (More-data-is-better-data.) A legtöbb közútról vannak precíz feljegyzések, digitalizálva. Akár több helyen is, amiből még biztosabb adatok szűrhetők le.

Ezeket megszerezve az adatokat tudják integrálni (emberi szaktudással „összegyűrti”). Minden új bizonyíték megerősíti a rendszert, javítja a logikai térképet. Bizonyos helyeknek fix helyzetük van, ami ismert, ezek a térképen elhelyezhetők, és a kapcsolódó objektumok hozzájuk igazíthatók. Konkrétan említik a közvilágítási oszlopok és a csatorna hozzáférések nyilvántartását, mint egy kapcsolódó forrást az utak pozíciójának keresésénél.

Ami érdekes a folyamatban, hogy ez az algoritmus nem ad tökéletes eredményt, ám a cég nem riadt vissza az eredmény emberi eszközökkel való javításától. Évek óta stabilan 6-700 fő dolgozik a térképen főállásban, minden apró hibát, egyenetlenséget kijavítva. Így konkrétan, kézzel beállítva a helyes asszociációkat sok helyen. Így tudnak olyan magas minőségű térképeket készíteni, mint például a Figure 9. ábrán látható.

Komoly a súlya a közösség segítségének (crowdsourcing) is. A felhasználók jó eséllyel kijavítják, bejelentik a saját maguk által tapasztalt hibákat. A cégeknek, üzleteknek érdekük, hogy megtalálhatóak legyenek, ezért aktívan szolgáltatnak adatot.



Figure 9: Adatok gyűrése

A jövőben újabb dimenziók is megnyílhatnak ezzel az értelmes adatszerkezéssel. Ha egy új tulajdonságot veszünk fel az alapigazságokhoz, akkor az lefordítható egyből földrajzi helyre. Így a rendszer magában rejti a lehetőségét, hogy egy igazán sokoldalú, hasznos térinformatikai rendszer legyen.

Például a Street view kamerás kocsikkal lefényképezte az összes utcát sok városban. Automatikusan fel tudják ismerni a kresz táblákat, és a pozíciójukat vissza tudják számítani. Egy béta változat eredménye látható a Figure 10. ábrán.



Figure 10: Egyéb objektumok megjelenítése

Itt azért érdemes megjegyezni, hogy azért lehetett ilyen látványosan gyorsan megoldani a problémát, mert sokat engedtek a precizitásból. Amíg egy tisztességes térinformatikai rendszernek mérnöki pontosságúnak kell lennie, jogi követelményeknek kell megfelelnie, itt erről nincs szó. Egy

„tisztességes” térinformatikai rendszerben általában felelősséget is vállalnak a szolgáltatott adatok minőségéért. A Google Maps viszont egy „best effort” szolgáltatást nyújt, díjmentesen, és ezt minden felhasználó tudja. A végfelhasználó a saját intelligenciájával kijavítja az egyes hibákat, pontatlanságokat és nem nagyon tudja felelősségre vonni az üzemeltetőket.

Google Data Center

Szerte a világon óriási szerverparkokban (Data Center) [5] futnak a Google szolgáltatásai. Összesen több, mint egy millió szerverük üzemel, egy-egy szerverpark több tízezer szervernek ad otthont. Ezeknek a szigorúan őrzött létesítményeknek a látványáról lehetnek elképzeléseink a Figure 11. ábra után.



Figure 11: Google Data Center

A Google stratégiája az, hogy nem csúcstechnológiájú szervereket alkalmaz, hanem közép árkategóriájú, kiforrott technológiájú hardverekből sokat. Optimalizálni elsősorban a skálázhatóságot és az üzemeltetési költséget szeretnék. Ezért is a saját fejlesztésű operációs rendszerük (Google File System) és adatbázis kezelőjük (Bigtable) fut az egyes szervereken elosztva.

Külön érdekesség, hogy olyan nagy tételben dolgoznak, hogy az egyes alkatrészeket nem is próbálják javítani, egyből cserélik. Olcsóbb a csere, mint a hibakeresés folyamata vagy egy feltáratlan lappangó hiba költsége. A rendszerben akkora redundancia van, hogy egy-egy alkatrész eltávolítása nem okoz fennakadást, és a rendszer újrakonfiguráció nélkül átveszi a feladatokat.

A gyakorlatban szó szerint arról van szó, hogy a karbantartó fizikailag kiveszi a hibás egységet, alkatrészt, szervert és betesz helyére egy újat. Minden feladatot real time a rendszer elvégez. A Figure 12. ábrán látható egy adag leselejtezett háttértár. Egy présgép fizikailag semmisítette meg ezeket adatvédelmi okokból.



Figure 12: Selejt alkatrészek

Néhány terhelési adat [4] a Google szolgáltatásokról:

- 20PB új adat érkezik hetente (Street view, műholdképek)
- 1 billió keresés, ami térkép asszociált is lehet
- 200TB felhasználói adat naponta csak a Google Mapshoz (24PB minden google szolgáltatás)
- Több mint egy millió szerver, cég összértéke 72.574 billió USD

Megjegyzhető itt is, hogy szerverterhelést és hálózati forgalmat minimalizálnak. Számításigényes műveleteket a felhasználókra hárítják (renderelés). Best effort szolgáltatást nyújtanak, amelyekben lehetnek üzemzavarok. A Google saját jelentése szerint 99%-os rendelkezésreállást tud biztosítani 2011-ben. (Hagyományos telekommunikációban elvárt a 99.9% vagy még jobb rendelkezésreállás is.)

Google Bigtable

A Google saját fejlesztésű adatbázis kezelő rendszere [7]. Specifikusan optimalizált a hardverre és a feladatra. Ezen tárolódnak az alapigazságok, a logikai térkép és maga a valódi térkép valamint az asszociációk. A konkrét adatbázis séma sajnos nem publikus.

A google Bigtable öbbségei:

- Gyors és extrém nagy méretű DBMS (PB nagyságrendű adatmennyiség, százezer gép)
- Ritka, elosztott, többdimenziós index (Oszlop- és sor orientált jellegzetességek is)
- Egyszerű több gépet hozzáadni. Nincs újrakonfigurálási költség.
- Több dimenziós táblák (verzió, időbélyeg)
- A Google File Systemre (GFS) optimalizálva
- A táblák töredékekre bontva tárolódnak (Sor alapján.) Kb. 200 MB egy töredék.

Szerver típusok:

- Master szerverek: Táblákat rendel a töredék szerverekhez. Számon tartja, hogy hol vannak a töredékek, és újraosztja ezeket, ha kell.
- Töredék szerverek: Írás, olvasás kiszolgálása. Tovább töri a táblát, ha a ráeső mérete meghaladja a limitet. Ugyancsak fontos, hogy ha kiesik, a rengeteg másik töredékszerver átveszi a feladatait automatikusan.
- Zár szerverek: A Chubby elosztott zárolási mechanizmus példányai.

Összefoglalás

Egy igen nehéz problémakörrel ismerkedhettünk meg, amelyre egyelőre nincs átütő erejű egységes megoldás. Mégis, ha a problémát a másik oldalról közelítjük meg, igen látványos és használható eredményeket kaphatunk. Ennek lényege, hogy nem az elkészített térképet akarjuk minél több információval ellátni (térinformatikai rendszer), hanem az információt strukturáljuk (a szemantikus háló által kínált technológiai elemek segítségével), és ezt rendeljük valós helyekhez a világban.

A technikát érdemes megismerni, mert a jövőben (ha nem is az egyetlen működő, de igen fontos) iránya lesz a fejlesztésnek. A Google módszerei napjainkban abszolút piacvezetőnek mondhatók, és megvalósításban is mindenképp a csúcstechnológiát jelentik. Érdemes a jövőben is figyelemmel kísérni az alkalmazott módszereiket.

Irodalomjegyzék

- [1] Michael Fowler: Time Dilation: A worked example (UVa Physics, 12/1/07)
http://galileoandstein.physics.virginia.edu/lectures/time_dil.html
- [2] Earth Measurement Consulting: GPS accuracy and limitations. 2000
http://earthmeasurement.com/GPS_accuracy.html
- [3] United States Census Bureau: TIGER 2010 <http://www.census.gov/geo/maps-data/data/tiger.html>
- [4] Alexis C. Madrigal: How Google Builds Its Maps—and What It Means for the Future of Everything (The Atlantic journal) 2012
<http://www.theatlantic.com/technology/archive/2012/09/how-google-builds-its-maps-and-what-it-means-for-the-future-of-everything/261913/>
- [5] Chris Burns: Google NC data center gets Street View treatment (Slashgear blog) 2012
<http://www.slashgear.com/google-data-center-hd-photos-hit-where-the-internet-lives-gallery-17252451/>
- [6] Wikipedia: Geographic information system 2012
http://en.wikipedia.org/wiki/Geographic_information_system
- [7] Wikipedia: Web mapping 2012 http://en.wikipedia.org/wiki/Web_mapping
- [8] World Wide Web Consortium: Semantic web 2012
<http://www.w3.org/standards/semanticweb/>
- [9] World Wide Web Consortium: RDF 2012
<http://www.w3.org/RDF/>