

Sijaintitiedon käyttö hakukoneissa

Juha Kuittinen

3.8.2006

Joensuun yliopisto
Tietojenkäsittelytiede
Pro gradu -tutkielma

Tiivistelmä

Internetsovelluksista suosituimpia ovat erilaiset *hakukoneet*. Tyypillisesti käyttäjä syöttää hakukoneeseen hakusanoja, joiden perusteella palautuu sanoja parhaiten vastaavien *www*-sivujen lista. Tässä tutkielmassa perehdytään mahdollisuuksiin hyödyntää käyttäjän sijaintitietoja yhtenä hakuena. Sijaintitiedon käytöllä pyritään järjestämään *www*-haun tuloksia siten, että niiden sisältämät kohteet sijaitsevat lähellä käyttäjää.

Sijaintiin perustuvan tiedon haku on hyödyllisintä liikkuvalla käyttäjälle. Siksi tutkielman yhteydessä on toteutettu sovellus, jonka avulla käyttäjä voi suorittaa itseään lähellä olevien palveluiden ja kohteiden hakuja mobiililla päätelaitteella. Käyttäjän paikannus tapahtuu erillisen GPS-laitteen avulla. Sovellus hyödyntää *www*-hakukoneen lisäksi myös kaupallisen palveluntarjoajan ilmaista palveluhakua.

Edellä mainitun sovelluksen avulla tutkitaan, voitaisiinko *www*-sivuja käyttää liikkuvan käyttäjän tietovarastona. Testitapausten avulla tutkitaan toteutetun sovelluksen käyttökelpoisuutta. Samalla käsitellään sekä sovelluksen suunnitteluun että yleisemmin *www*-hakuihin liittyviä ongelmia.

ACM-luokat (ACM Computing Classification System, 1998 version): H.3.3, H.3.5

Avainsanat: hakukone, lbs, mobiili päätelaite, mobiili paikannus, paikkatieto, sijaintitieto, *www*.

Esipuhe

Haluan lausua kiitokseni professori Pasi Fräntille ohjauksesta sekä loistavasta aiheesta joka vei mennessään. Erityiskiitos myös tohtori Alexander Kolesnikoville. Hän räätälöi tekemänsä koordinaattien muunto-ohjelman sovellukseni käyttöön.

Sisältö

1	Johdanto	1
2	LBS-sovellukset	3
3	Koordinaateista ja paikannusmenetelmistä	5
3.1	Koordinaattijärjestelmät	5
3.2	Paikannusmenetelmiä	7
4	Paikkatiedon esiintyminen www -dokumenteissa	12
5	Sijaintitiedon hyödyntäminen hakukoneissa	16
5.1	Käyttäjän sijainnin määrittäminen	17
5.2	Tietolähteet	17
5.2.1	Tiedonhaku www-sivuilta	18
5.2.2	Tiedonhaku tunnetuilta palveluntarjoajilta	19
5.3	Relevantin tiedon eristäminen tietolähteistä	19
5.4	Tiedon prosessointi päätelaitteelle	20
6	Koesovelluksen toteutus käytännössä	21
6.1	Ohjelmointitekniikat ja laitteisto	22
6.2	Hakuprofiili	24
6.3	Kohteiden haku verkosta	24
6.3.1	Www-haku Googlella	25
6.3.2	Haku Fonectan tietokannasta	26
6.4	Paikkatiedon eristäminen	27
6.5	Ohjelman käyttämät tietovarastot	30
7	Koesovelluksen käyttökokemuksia	34
7.1	Testituloksia	35
7.1.1	Matkailija vieraassa kaupungissa	36
7.1.2	Matkailija liikkeellä maaseudulla	37
7.1.3	Hakuprofiilitestin tuloksia	40
7.1.4	Www-hakujen saanti ja tarkkuus	42
7.2	Havaittuja ongelmia	44
7.2.1	Sijaintitiedon eristämisoongelmat	45
7.2.2	Osoitteet ja koordinaatit	46

7.2.3	Hakutuloksiin liittyviä ongelmia	46
7.2.4	Sovelluksen käytettävyyteen ja testilaitteistoon liittyviä ongelmia	47
7.2.5	Muita ongelmia	48
8	Kehitysideoita	51
8.1	Tiedonhaun tehostaminen	51
8.2	Sijaintitiedon eristäminen www-sivuilta	51
8.3	Hakutulosten laatu ja määrä	52
8.4	Palvelimella sijaitsevat tietovarastot	53
8.5	Käyttäjän sijaintikunnan määrittämisongelma	54
8.6	Maariippumattomuus	56
9	Hakutuloksien parantamismahdollisuuksia	57
9.1	Nykymuotoinen web	57
9.2	Semanttinen web	59
10	Yhteenveto	63
	Viitteet	65

1 Johdanto

Missä on lähin pizzeria? Jos kysyjä tietää sijaintinsa ja hänellä on käytössään internetiyhteydellä varustettu tietokone, asiaa voidaan yrittää selvittää www-haun avulla. Kysyjä voi käyttää tavallista hakukonetta, johon hän syöttää ”pizzeria”-hakusanan lisäksi paikannimen. Hakutuloksista saattaa löytyä kuitenkin myös paljon käyttäjän kannalta hyödyttömiä tai muuten hankalakäyttöisiä sivuja. Erilaiset palvelusivustot (esimerkiksi www.020202.fi, www.keltaisetsivut.fi) tarjoavat mahdollisuuden hakea tietoa käyttäen yhtenä hakukriteerinä paikkakuntaa tai aluetta. Onnistuessaan haut antavat tarkkoja sijaintitietoja ja kertovat tarvittaessa myös reitin perille. Edellytyksenä kuitenkin on, että yritys tai palvelu on ostanut itselleen oikeuden näkyä palvelusivuston käyttämissä tietokannoissa.

Www:n rakenne ja nykyiset www-hakukoneet eivät tarjoa tukea sijainnin käyttämiselle hakuehtona. Hakusanana on mahdollista käyttää paikkakunnan nimeä, mutta hakutuloksissa on runsaasti myös sivuja jotka sisältävät kyseisen sanan jossakin muussa merkityksessä. Toisaalta www-sivujen tuottajat eivät juurikaan käytä sijaintiedon esittämiseen siihen tarkoitettuja standardeja, joten tiedon löytäminen www-sivuilta ei onnistu luotettavasti. Käyttäjän sijaintia kuvaavaa paikkannimeä voidaan kuitenkin käyttää hakua rajaavana hakusanana. Näin voidaan lisätä todennäköisyyttä, että haku palauttaa sijaintitietoa sisältäviä www-sivuja.

Sijaintitiedon käyttö hakuehtona on hyödyllisintä käyttäjälle joka on liikkeellä, eikä tiedä tarkasti omaa sijaintiaan. Erilaiset mobiilit päätelaitteet, kuten esimerkiksi älypuhelimet, sisältävät nykyisin mahdollisuuden selailla www-sivujen sisältöä. Tavallisten www-hakukoneiden palauttavat tulossivut ovat kuitenkin usein liian raskaita ja sisältävät liikaa informaatiota mobiililaitteita ajatellen. Lisäksi tuloksena saadut sivut eivät suoraan kerro käyttäjälle *lähimmän* kohteen sijaintia, vaan hänen on pääteltävä se itse osoitetietojen tai kartan perusteella. Tässä tutkielmassa on tarkoitus kartoittaa mobiilin, sijaintitietoa hyväksikäyttävän hakukoneen toteuttamiseen liittyviä ratkaisuja ja ongelmia.

Luvussa 2 esitellään aihetta lähellä olevaa LBS (Location-Based Services) -sovellusalueutta. Sen jälkeen luodaan katsaus sijaintitiedon esittämisen ja paikantamisen perusteisiin. Luvussa 4 käydään läpi paikkatiedon esitystapoja www-dokumenteissa. Luvussa 5 tutkitaan sijaintiin liittyvään tiedonhakuun kykenevän mobiilin sovelluksen

toteuttamiseen liittyviä ideoita ja näkökohtia. Luvussa 6 esitellään yksityiskohtaisemmin tämän tutkielman yhteydessä toteutettu koesovellus. Koesovellus on mobiililaitteella käytettävä sovellus, jonka avulla tutkitaan, voidaanko www-hakuja käyttämällä saada riittävän tarkkoja ja käyttäjän sijainnin suhteen relevantteja tuloksia. Vertailun vuoksi sovellus suorittaa samat haut myös kaupalliselle palvelusivustolle. Koesovellus toimii ainoastaan Suomen alueella ja käyttää tietovarastoinaan suomalaisia www-sivuja. Luvussa 7 testataan sovellusta sekä tuodaan esille toteutuksessa havaittuja ongelmia. Ongelmien mahdollisia ratkaisuja ja jatkokehitysnäkymiä käydään läpi luvussa 8.

Www-dokumenttien muodostaman valtavan tietovaraston tehokas hyödyntäminen on nykyisillä hakumenetelmillä vaikeaa. Luvussa 9 käsitellään www-hakuihin liittyviä ongelmia yleisemmällä tasolla sekä tutustutaan www:n seuraavaan mahdolliseen kehityssaskeleeseen, semanttiseen webiin.

2 LBS-sovellukset

Mobiiliteknologia mahdollistaa nykyisin reaaliaikaisen paikannuksen ja internetsovellusten käytön myös käyttäjän ollessa liikkeellä. Tämä kehitys on avannut tietä kasvavalle joukolle paikannukseen perustuvia palveluita. *Paikannukseen perustuvat palvelut (LBS-palvelut, Location-Based Services)* ovat käyttäjän sijaintitietoa hyödyntäviä, mobiililaitteilla käytettäviä palveluita [25]. Tulevaisuudessa suurimman osan mobiilipalveluista ennustetaan olevan LBS-palveluita [25].

LBS-palveluihin perustuva liiketoiminta on vielä kehitysvaiheessa ja markkinoilla on tällä hetkellä muutamia sovelluksia. Esimerkkinä voidaan mainita Soneran ”*Missä olet*” -palvelu, jossa henkilöitä voidaan paikantaa heidän matkapuhelimiensa sijainteja tutkimalla [22]. Paikannukseen luvan antaneen henkilön sijainti määritetään matkapuhelinverkon avulla ja palautetaan kysyjälle esimerkiksi karttapisteenä. Samalla palvelulla käyttäjä voi selvittää myös oman sijaintinsa kartalla. Toinen saman yhtiön palvelu tarjoaa matkapuhelimen sijainnin perusteella tietoa lähiympäristössä sijaitsevista palveluista.

LBS-palvelut voidaan jakaa toiminnallisuuden ja paikkatiedon hyödyntämistapojen perusteella eri luokkiin [25]. Tyypillisin LBS-palveluiden ryhmä tuottaa käyttäjän sijaintiin perustuvaa informaatiota. Käyttäjän sijaintia lähellä olevien kohteiden etsintään tarkoitettut palvelut kuuluvat tähän luokkaan.

Toinen LBS-palveluiden ryhmä tuottaa tiedon sijasta käyttäjän sijaintiin liittyviä palveluita. Eräs esimerkki tällaisesta voisi olla taksitilauspalvelu. Kuvitteellisen ”*Tilaa taksi*” -napin painalluksella käyttäjän sijainti välittyy palvelulle. Palvelu ohjaa tämän jälkeen automaattisesti paikalle lähimmän vapaan taksin. Toisena esimerkkinä voitaisiin mainita erilaiset hälytys- ja turvallisuuspalvelut. Tällaisia sovelluksia on jo käytössä esimerkiksi vanhustenhoidossa.

Omana ryhmänään ovat palvelut, joissa käyttäjän oman sijainnin sijasta tutkitaan toisten käyttäjien sijainteja (jäljityspalvelu). Esimerkkinä tällaisista sovelluksista on aiemmin mainittu Soneran ”*Missä olet*” -palvelu.

Edellä kuvatut palvelut vaativat käyttäjältä omaa aktiivisuutta. Tämän tyyppisiä palveluita kutsutaan LBS-kentässä ”*pull*”-palveluiksi [25]. Sijaintiin perustuvia palveluita voidaan tarjota myös käyttäjän sitä erikseen pyytämättä. Tähän LBS-ryhmään kuulu-

vat palvelut voivat esimerkiksi lähettää käyttäjälle hänen tämänhetkisen sijaintinsa läheisyydessä olevien yritysten mainoksia. Näitä palveluita sanotaan ”*push*”-palveluiksi.

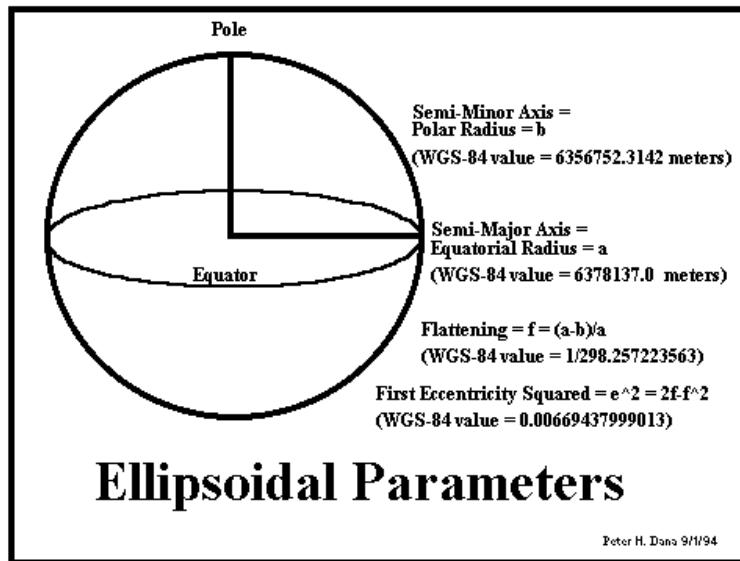
Tämän tutkielman ohessa toteutettu koesovellus etsii käyttäjän pyynnöstä lähellä olevia kohteita ja palveluja. Tämän eräänlaisen internet-hakukoneen ja LBS-sovelluksen yhdistelmän voidaan katsoa kuuluvan yleisimpään, sijaintiin liittyvää informaatiota tarjoavien LBS-palveluiden ryhmään.

3 Koordinaateista ja paikannusmenetelmistä

Paikkatieto tarkoittaa paikannettua kohdetta kuvaavan sijaintitiedon ja kohteen ominaisuuksia kuvaavien tietojen muodostamaa kokonaisuutta. *Sijaintitieto* on puolestaan vertausjärjestelmässä ilmaistu kohteen sijainti. *Vertausjärjestelmä* on järjestelmä, jonka avulla sijainti voidaan ilmaista yksikäsitteisesti. Vertausjärjestelmänä voi olla esimerkiksi jokin koordinaattijärjestelmä tai osoitejärjestelmä [24]. Maapallolla on käytössä useita eri vertausellipsoideihin perustuvia koordinaattijärjestelmiä.

3.1 Koordinaattijärjestelmät

Maa ei ole pallo, vaan navoiltaan litistynyt. Tämän kuvaamiseksi käytetään *vertausellipsoideja*, jotka ovat geometrisesti ajatellen ellipsin pyörähdyskappaleita. Maan ajatellaan olevan ellipsin muotoinen kappale, joka pyörii lyhemmän puoliakselinsa ympäri [4]. Vertausellipsoideja on ollut käytössä useita. Erot johtuvat mittaustavoista sekä siitä alueesta, johon elipsoideja on yritetty sovittaa. Erot näkyvät eniten *lyhemmän- ja pitemmän puoliakselin* (semi-minor-, semi-major axis) mitoissa sekä *litistyneisyydessä* (flattening).



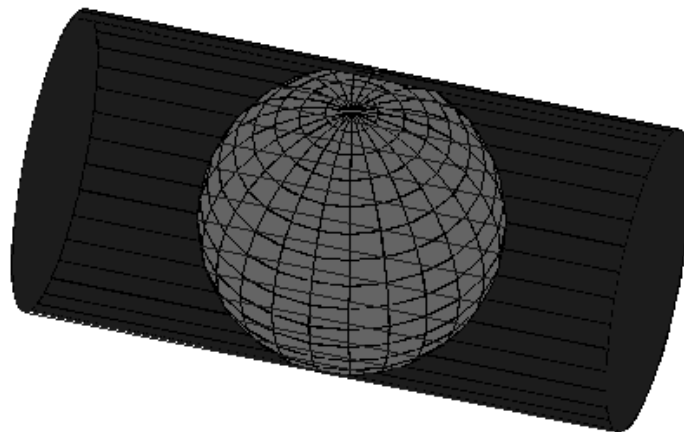
Kuva 1: Vertausellipsoidi [2].

Koordinaattijärjestelmiä (Datum) luotaessa käytetään apuna vertausellipsoidista saatuja tietoja, yleensä pitemmän puoliakselin pituutta sekä litistyneisyyttä. Koska ellipsoi-

di on muodoltaan säännöllinen, se ei vastaa riittävästi maanpinnan monimuotoisuutta. Jotta koordinaatiston määrittäminen onnistuisi riittävällä tarkkuudella, on tehtävä korjaavia muutoksia. Lisäksi on kiinnitettävä pohjoisnapa, keskipiste sekä nollameridiaani. Näiden toimien tuloksena syntyvät koordinaattijärjestelmät [24]. GPS-käytössä koordinaattijärjestelmänä on yleensä WGS84 (World Geodetic System). Järjestelmä sisältää maantieteellisen koordinaatiston (ellipsoidikoordinaatiston), joten se tuottaa koordinaattipisteet asteina, minuutteina ja sekunteina.

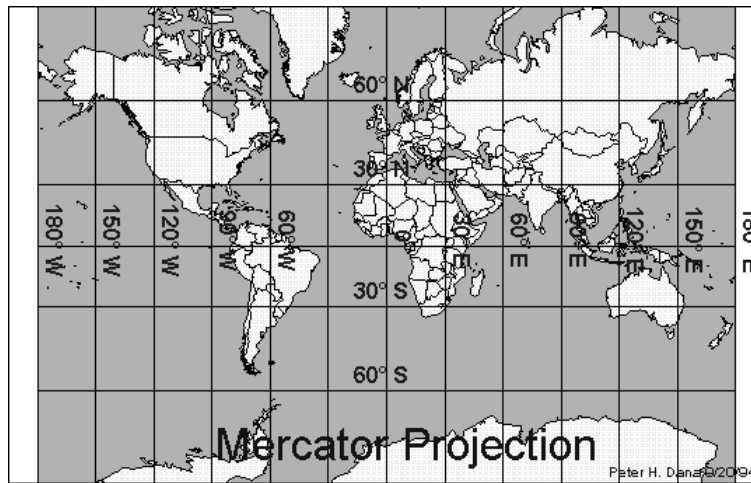
Koordinaatistossa pisteen sijainti ilmoitetaan lukuina tai kulmina (*koordinaatteina*) valitussa koordinaattijärjestelmässä. Tavallisia suomalaisia maastokarttoja varten täytyy koordinaattijärjestelmän maantieteelliset koordinaatit muuttaa tasokoordinaateiksi. Koska ellipsin pintaa ei voida suoraan siirtää tasolle, on käytettävä *karttaprojektioita*. Suomalaisissa kkj (kansallinen kartastojärjestelmä) perus- ja yhtenäiskoordinaatistoissa käytetään Gauss-Krügerin poikittaista sylinteriprojektiota. Tässä tapauksessa tasopinta ”kääritään” pallopinnan ympärille ja tallennetaan pallopinnan koordinaatiston projektio tasolla [3]. Tarkkuus on hyvä ainoastaan silloin kun ”kääre” koskettaa pallopintaa, joten projektioita on otettava useita, tasoa aina välillä kääntäen. Näin muodostuu projektiokaistoja, joista yhdistämällä saadaan karttaprojektio. Kuvassa 2 on havainnollistettu sylinterimäisen *Transverse* projektion muodostumista. Kuvassa 3 puolestaan valmis *Mercator* karttaprojektio.

Peter H. Dana 10/01/94



Transverse Cylindrical Projection Surface

Kuva 2: Transverse -sylinteriprojektion periaate [3].



Kuva 3: Mercator-karttaprojektio [3].

Suomalaisessa *kkj*-koordinaatistossa koordinaatit ovat lukuja, jotka kuvaavat etäisyyttä päiväntasaajalta pohjoiseen sekä projektiokaistan etäisyyttä keskimeridiaanilta itä-länsi suunnassa. Maanmittauslaitoksen ylläpitämissä kartoissa ja tietokannoissa on vielä käytössä edellämainittu kansallinen koordinaatisto. Sijaintia hakehtona käyttävän sovelluksen tekijän onkin otettava huomioon erilaiset kansalliset koordinaatistot. Esimerkiksi tämän tutkielman ohessa toteutetun koesovelluksen osoite/koordinaatti -tietokantana toimii maanmittauslaitoksen ”*Kansalaisen karttapaiikka*” -palvelu [14]. Sovellus saa koordinaatit osoitteista *kkj*-yhtenäiskoordinaatisto -muodossa. Koska käyttäjän sijainti saadaan GPS-paikantimen maantieteellisinä koordinaatteina (WGS84), joudutaan koesovelluksessa suorittamaan koordinaattimuunnoksia.

Maanmittauslaitos on piakkoin siirtymässä *ETRS89* (European Terrestrial Reference System) koordinaattijärjestelmään perustuvaan *EUREF-FIN* koordinaatistoon [15]. Järjestelmän on todettu yhtyvän WGS84-järjestelmään noin metrin tarkkuudella. Tämän suuntaan kehitys helpottaa esimerkiksi juuri koesovelluksen kaltaisen ohjelmiston kehittämistä, sillä koordinaattien välisten muunnoksien tarve vähenee.

3.2 Paikannusmenetelmiä

Paikannusjärjestelmä tarkoittaa sijaintitiedon määrittelyä järjestelmää [24]. Mobiilit paikannusjärjestelmät perustuvat tyypillisesti etäisyyksien mittaukseen signaalin kuluaikaa havainnoimalla [20]. Paikannusjärjestelmät voidaan jakaa

toimintaperiaatteensa mukaisesti kolmeen ryhmään:

1. Verkkopaikannusjärjestelmät.
2. Lähipaikannusjärjestelmät.
3. Satelliittipaikannusjärjestelmät.

Verkkopaikannuksella tarkoitetaan matkaviestinverkon (esimerkiksi GSM) avulla tapahtuvaa paikannusta. Paikannuksen mahdollistavia tekniikoita on useita erilaisia. Tekniikoiden paikannustarkkuus ja käyttökelpoisuus eri alueilla ja eri tilanteissa on vaihteleva. Yksinkertaisin menetelmä on *solupaikannus*, jossa verkkoon kytkettynä olevan puhelimen sijainti voidaan arvioida puhelimeen yhteydessä olevan tukiaseman perusteella. Tukiaseman peittoalueen eli *solun* tunnisteella voidaan tunnistaa tukiasema. Paikannuksen tarkkuus vaihtelee solun koosta riippuen 200 metristä jopa 25 kilometriin [18, 20]. Solupaikannus oli vuonna 2005 yleisin käytössä oleva verkkopaikannusmenetelmä. Tunnetuimmat verkkopaikannusmenetelmät ovat [20]:

- *Solupaikannus* (CI, Cell Identification).
- *Solupaikannus täydennettynä ajastusennakko- ja signaalin voimakkuustiedoilla* (CI+TA+Rx; TA, Timing Advance; Rx, signal level).
- *Saapumiskulmapaikannus* (AOA, Angle of Arrival).
- *Saapumisaikapaikannus* (TOA, Time of Arrival).
- *Kulkuaikaeropaikannus* (EOTD, Enhanced Observed Time Difference).
- *Saapumisaikaeropaikannus* (TDOA, Time Difference of Arrival).
- *Korrelaatiopaikannus* (LF, Location Fingerprint).

Lähipaikannuksella tarkoitetaan sisätiloissa tai esimerkiksi kaupunkien keskustoissa tapahtuvaa langatonta tarkkuuspaikannusta. Lähipaikannuksen menetelmiä ovat esimerkiksi [20]:

- *Lähiverkkopaikannus*.
- *Bluetooth-paikannus*.

- *Etätunnistipaikannus.*
- *Infrapunapaikannus.*
- *Ultraäänipaikannus.*

Rakennusten sisällä tapahtuva paikannus voisi olla hyödyllistä esimerkiksi suurissa julkisissa tiloissa, mutta tämän tutkielman kannalta lähipaikannusmenetelmistä kiinnostavin on lähiverkkopaikannus. GPS-paikantimet eivät yleensä toimi sisätiloissa ja verkkopaikannuksen tarkkuus ei välttämättä ole riittävä kaupunkikäyttöä ajatellen. Balachandran & al.(2003) [21] esittelevät artikkelissaan langattomien verkkojen liityntäpisteisiin (eng. *hotspot*) perustuvaa päätelaitteen paikannusta (*WiFi Positioning*). Käyttäjän avatessa langattoman päätelaitteensa, sen sijainti määritetään yhteysetäisyydellä olevien verkon liityntäpisteiden avulla. USA:n suurissa keskuksissa aletaan jo olla tilanteessa, jossa hotspotien tiheys mahdollistaa GPS-tasoisin paikannuksen [21]. To- teutuakseen tämänkaltaisen paikannus vaatisi käyttäjän pääsyn ajantasaistettuun, kaik- ki hotspotit sisältävään tietokantaan.

Satelliittipaikannus perustuu paikannussatelliittien lähettämän ratatieto- ja aikamerkki- signaalin vastaanottoon ja vastaanottimen sijainnin laskemiseen satelliittien etäisyyden perusteella [20]. Yleisessä käytössä on Yhdysvaltain puolustushallinnon ylläpitämä GPS (Global Positioning System) satelliittipaikannusjärjestelmä. Järjestelmään kuu- luu vähintään 24 maata kiertävää satelliittia. Kohteen paikannukseen tarvitaan yleensä vähintään neljän satelliitin signaali. Järjestelmä käyttää lisäksi maa-asemia, jotka muun muassa laskevat satelliittien ratatietoja. Satelliitin etäisyys vastaanottimesta lasketaan signaalin saapumisajan perusteella. Ratatietojen perusteella tiedetään satelliitin paik- ka tietyllä ajanhetkellä. Kun nämä tiedot yhdistetään, voidaan vastaanottimessa laskea senhetkinen sijainti. Myös satelliittipaikannuksessa voidaan käyttää erilaisia menetel- miä. Seuraavassa on listattu muutamia [20]:

- *GPS-paikannus* (GPS, Global Positioning System).
- *Differentiaalinen GPS* (DGPS, Differential GPS).
- *Reaaliaikainen kinemaattinen GPS* (RTK GPS, Real Time Kinematic GPS).
- *GPS-pseudoliittipaikannus.*
- *Avustettu GPS-paikannus* (AGPS, Assisted Global Positioning System).

Differentiaalisessa GPS-paikannuksessa paikallisten mittausasemien lähettämää korjausdataa käytetään paikannustarkkuuden parantamiseen. Vaihtoehtona paikallisille mittausasemille voivat olla maan suhteen paikallaan pysyvät satelliitit. *Reaaliaikainen kinemaattinen GPS* on lähinnä ammattimaisten paikannustiedon käyttäjien (esim. maanmittaus) menetelmä, jossa yksi mittausyksikkö on ennestään tunnetussa pisteessä. Paikannusta voidaan suorittaa muutaman kilometrin säteellä mittausyksiköstä.

GPS-pseudoliittipaikannus perustuu *pseudoliittien* eli valesatelliittien lähettämiin signaaleihin [9]. Pseudoliitit ovat GPS-signaaleja lähettäviä laitteita, jotka on asennettu paikkoihin missä oikeaa satelliitti-signaalia ei pystytä vastaanottamaan. Näin menetellen voidaan esimerkiksi sisätiloihin rakentaa ”satelliittiverkko”, jonka signaaleja voi ottaa vastaan GPS-vastaanottimella. Ulkotiloihin siirryttäessä vastaanotin siirtyy käyttämään taas tavallista satelliittipaikannusta.

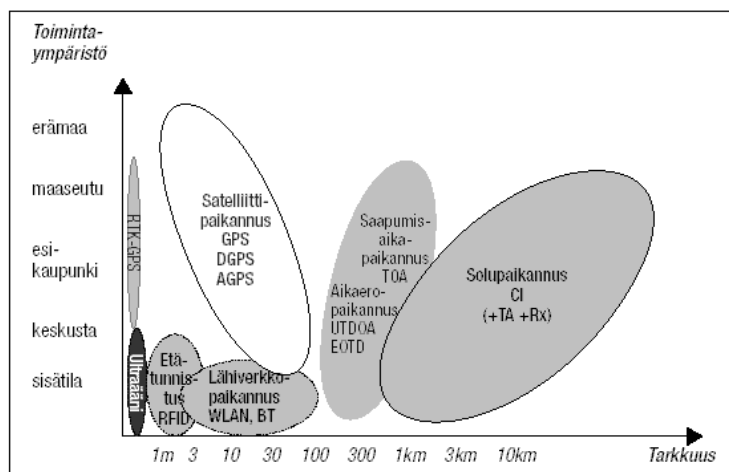
Avustettu GPS-paikannus yrittää myös löytää ratkaisua GPS-paikannuksen ongelmiin satelliittisignaalin ollessa heikko. Toinen tavoite on vähentää vastaanottimen virrankulutusta [20]. Standardin mukaisia toimintaperiaatteita on kahdenlaisia. Ensimmäisessä matkapuhelinverkko välittää vastaanottimelle päivitetyt rata- ja korjaustiedot. Edellytyksenä on matkaviestimen ja satelliittivastaanottimen yhteistoiminta. Toisessa vaihtoehdossa vastaanotin lähettää näytteen satelliittisignaalista verkon palvelimelle, joka huolehtii sijainnin laskemisesta. Avustettu satelliittipaikannus on jo käytössä laajasti USA:ssa, Japanissa ja Etelä-Koreassa [20]. Avustettua GPS-paikannustekniikkaa hyödyntävien AGPS-puhelinten myynnin ennustetaan olevan Euroopassa vuonna 2007 yli 50 miljoonaa.

GPS ei ole ainoa satelliittipaikannusjärjestelmä. Venäjällä on oma *Glonass* ja Eurooppa on rakentamassa *Galileo*-järjestelmäänsä. Suunnitelmien mukaan Galileo koostuisi 30:stä satelliitista, joista 27 olisi aktiivisena [20]. Järjestelmässä olisi kolmen tasoisia palveluita: kaikille ilmainen avoin palvelu, kaupallinen palvelu sekä viranomaispalvelu. Näistä kaupallinen tarjoaa kaikkein tarkinta paikannusta. On myös mahdollista, että tulevaisuudessa GPS- ja Galileo-järjestelmiä voitaisiin käyttää yhtäaikaaisesti samalla vastaanottimella. Tällöin tarkkuus voisi parantua, koska käytettävissä on samanaikaisesti enemmän satelliitteja. Vuonna 2005 laukaistiin radalleen ensimmäinen Galileo testisatelliitti. Aikataulun mukaan järjestelmän tulisi olla toiminnassa vuonna 2008.

Tässä luvussa käsitellyillä langattomilla paikannusmenetelmillä on kaikilla omat vahvuutensa ja heikkoutensa. GPS on tarkka suunnistusapu, mutta se ei ulotu sisätiloihin.

Sisätiloissa käytettävät menetelmät eivät puolestaan sovi karttanavigointiin. Paras tulos olisi saavutettavissa yhdistelemällä edellä kuvattuja erilaisia paikannusmenetelmiä samalla laitteistolla käytettäväksi. Tällaista eri menetelmien yhdistämistä kutsutaan hybridipaikannukseksi [20].

Erilaisten langattomien paikannusmenetelmien tarkkuutta ja käyttökohteita on havainnollistettu kuvassa 4. Kuvaajan vaaka-akselilla on kuvattu paikannustarkkuutta ja pystyakselilla puolestaan eri menetelmien käyttökelpoisia toimintaympäristöjä. Kuvasta voidaan havaita, että parasta tarkkuutta (< 1 metri) tarjoavat lähipaikannusmenetelmiin kuuluva ultraäänipaikannus sekä satelliittipaikannukseen kuuluva reaaliaikainen kinemaattinen GPS (RTK-GPS). Tämä menetelmä tosin vaatii ennalta määritettyjen mittauspisteiden läheisyydessä pysyttelemistä, joten se ei sovellu vapaasti liikkuvien kohteiden paikantamiseen. Jos tarvitaan liikkuvan kohteen paikannusta rakennusten ulkopuolella, parhaan tarkkuuden ja toiminta-alueen yhdistelmän tarjoaa satelliittipaikannus.



Kuva 4: Paikannusmenetelmien vertailua [20].

4 Paikkatiedon esiintyminen www -dokumenteissa

WWW (World Wide Web) sai alkunsa CERN-laboratorioilla Genevessä [1]. Pyrkimyksenä oli luoda tietolähde, joka mahdollisti eri paikoissa sijaitsevien toimijoiden jakaa ideoitaan ja näkökulmiaan toistensa kanssa. Tutkijoilla oli tähän tarkoitukseen tehty pieni hypertekstisovellus. Positiiviset kokemukset panivat alulle ajatuksen World Wide Webistä ja loivat pohjan edelleenkin käytössä oleville HTTP- ja HTML -standardeille. Edellä mainittuja sekä monia muita www -standardeja kehittää ja valvoo W3C (World Wide Web Consortium) -organisaatio [28].

Hyperteksti sisältää linkkejä, joiden avulla tekstistä voi viitata toisiin dokumentteihin tai vaikkapa liikkua tekstissä vapaasti eri suuntiin. Linkkien avulla tekstiin voi lisätä myös ääntä ja kuvaa. Tällöin puhutaan *hypermediasta*.

HTML (Hypertext Markup Language) on erityisesti hypertekstin välittämiseen suunniteltu kieli. Kieli sisältää yksinkertaisia rakenne-elementtejä kuten otsikoita, listoja, luetteloita ja niin edelleen. Elementtien avulla määrittyy, kuinka tietyt tekstin osat näkyvät selaimissa. Esimerkiksi otsikkoelementiksi merkitty teksti eroaa selvästi tavallisesta tekstistä. Elementtejä merkitään *tageilla*. Tagit erottuvat tekstistä `<>` ja `</>` merkeillä.

Html-dokumentissa on mahdollista esittää myös *metatietoa*, jonka avulla tekijä voi antaa tietoa dokumentin sisällöstä. Tieto esitetään meta-tagien sisällä ominaisuus-arvo pareina. Tekijä voi esimerkiksi kertoa valokuvausta käsittelevän dokumenttinsa sisällöstä lisäämällä rivin:

```
<meta name="keywords" content="valokuvaus, kamera">
```

Ominaisuutta on kuvattu esimerkissä merkkijonolla "keywords". Usein käytettyjä ominaisuuksia meta-tageissa ovat esimerkiksi "keywords", "author" ja "description". Html-standardi ei määrittele sallittujen ominaisuuksien joukkoa, joten sisältöä on mahdollista kuvailla myös itse keksityillä ominaisuuksilla. Jotta kuvaus tulisi noteeratuksi esimerkiksi hakukoneissa, sen tulisi noudattaa totuttuja ja sovittuja käytäntöjä.

Jotta www-sivulta haettava kohde voitaisiin paikallistaa, pitää html-koodista kyetä eristämään sijaintitietoa. Www-sivuilla esiintyvää sijainti- ja paikkatietoa on tutkittu jonkin verran [16, 26]. Ilkka Vänskä tutki pro-gradu tutkielmassaan 24000 suomalaista

www-sivua etsien niistä erilaisia paikkatietoelementtejä [26]. Edellä mainittujen lähteiden perusteella sijainti- ja paikkatietoa esiintyy www-dokumenteissa seuraavilla tavoilla:

- Geotagit.
- Address-tagit.
- Osoitteet.
- Postinumerot.
- Puhelinnumerot.
- Maantieteelliset kohteet.

Geotagien (geotags) avulla voidaan metatietoon kirjoittaa koordinaattien lisäksi myös tiedot maasta ja alueesta [6]. Esimerkkinä tagien käytöstä on seuraavassa Joensuun tiedepuiston sijainti.

```
<meta name="geo.position" content="62.598;29.744" />
<meta name="geo.placename" content="Joensuu, Finland" />
<meta name="geo.region" content="FI-IS" />
<meta name="geo.country" content="FI" />
```

Koordinaatit kirjoitetaan ”geo.position” -kenttään maantieteellisinä desimaaliasteina. Koordinaatit kirjoitetaan järjestyksessä leveys ja pituus puolipisteellä erotettuna. Positiiviset luvut tarkoittavat pohjoista ja itäistä, negatiiviset eteläistä ja läntistä. Vertausjärjestelmänä WGS84 antaa riittävän tarkkuuden. ”Geo.placename” -kenttään voidaan kirjoittaa vapaamuotoista tekstiä, mutta geotags.com:in suositus on tukeutua standardeituihin paikannimiin, esimerkiksi ”*Getty Thesaurus of Geographic Names*”¹ tietokannan nimistöön. ”Geo.region” sisältää ISO-3166-2-standardin mukaisen aluekoodin, ”geo.country” puolestaan ISO-3166-1:ssä määritellyn maakoodin. Standardien huolellinen käyttö on edellytys sille, että paikkatiedon löytäminen www-dokumenteista voidaan toteuttaa tehokkaasti ja maariippumattomasti.

¹<http://www.getty.edu/research/tools/vocabulary/tgn/>

Geotags.com pitää yllä hakukonetta, joka etsii geotageilla varustettuja www-dokumentteja. Hakukoneeseen syötetään paikannimi, jonka jälkeen hakukone etsii kyseisen paikan sisältäviä dokumentteja. Tutkielman kirjoittamisen aikoihin hakukone löysi esimerkiksi ”Helsinki”-hakusanalla vain kaksi geotageilla varustettua www-sivustoa. Muutamilla muilla isohkojen kaupunkien nimillä ei löytynyt yhtään sivustoa. Ilmeisesti geotagien käyttö on ainakin toistaiseksi hyvin vähäistä.

ADDRESS-tagi on html-standardin hyväksymä tapa lisätä sijainti omana elementtinään [28]. Tieto lisätään dokumenttiin kirjoittamalla osoite `<ADDRESS>` ja `</ADDRESS>` tagien väliin. Web-selaimet tunnistavat tagin ja esittävät annetut sijainti- ja yhteystiedot kursivoituna. Html-tuesta huolimatta tagin käyttö ei ole kovin yleistä [26].

Koska sijaintitiedon määrittely tagien avulla on varsin harvinaista, joudutaan useimmiten tutkimaan itse sivun tekstiä. Paras tulos saadaan, jos tekstistä onnistutaan löytämään kokonainen osoite. Osoite voidaan muuntaa sijaintikoordinaateiksi käyttämällä jotakin sopivaa osoite/koordinaatti -tietokantaa. Esimerkiksi Suomessa maanmittauslaitos pitää yllä verkossa käytettävää tietokantaa ², josta katuosoitteen ja kunnan nimen perusteella saadaan osoitteen karttakoordinaatit. Maakohtaisia osoite/koordinaatti -tietokantoja on käytössä useissa kehittyneissä maissa, esimerkiksi USA:ssa *Tiger/Zip+4* [16].

Sijainteja voi päätellä myös pelkän postinumeron perusteella. Suomessa suurin osa kunnista sisältää useita postinumeroalueita. Alueiden koko kuitenkin vaihtelee eri osissa maata, mikä aiheuttaa paikannustarkkuuden vaihtelua. Eri maat noudattavat muoltoltaan erilaisia postinumerostandardeja, joten numeroita tekstistä etsiessä pitäisi tietää, minkä maan paikkatietoa haetaan. Kaikki postinumeron muotoiset numerot eivät myöskään aina ole postinumeroita, joten virhetulkintoja voi tapahtua.

Puhelinnumeron perusteella on parhaassa tapauksessa mahdollista päätellä maa ja/tai alue. Usein numeroita ei kuitenkaan kirjoiteta täydellisinä: maakoodi saattaa puuttua tai aluekoodia ei ole katsottu tarpeelliseksi käyttää. Vakiomuotoiset numeroesitykset saattavat myös erota maakohtaisesti [16]. Lisäksi kirjoittaja voi ottaa vapauksia numeron esitystavoissa lisäämällä esimerkiksi ylimääräisiä sulkuja tai välilyöntejä. Puhelinnumeron eristäminen tekstistä ei siten ole täysin luotettavaa.

Muita paikkaan liittyviä ilmauksia ovat esimerkiksi maantieteellisten kohteiden nimet.

²<http://kansalaisen.karttapaikka.fi>

Näitä voivat olla järvet, joet, vuoret tai vaikkapa tunnetut rakennukset. Sijainti voidaan selvittää käyttämällä kyseisiä kohteita sisältäviä tietokantoja. USA:n alueella eräs tällainen on *Geographic Names Information System* (GNIS) [16].

USA:laisen tutkimuksen mukaan 4.5% kaikista www-sivuista sisälsi tunnistettavan postinumeron, 8.5% puhelinnumeron ja 9.5% jomman kumman edellisistä [16]. Ilkka Vänskä puolestaan on pro-gradu tutkielmassaan [26] tutkinut 24000 suomalaista www-sivustoa. Vastaavat prosenttiluvut hänen materiaalissaan olivat 16.4%, 14.8% ja 20.5%. Ensiksimainitussa tutkimuksessa mukaan oli laskettu vain varmat tapaukset ja jälkimmäisessä taas kaikki muotonsa puolesta paikkatiedoksi kelpaavat tapaukset (esimerkiksi kaikki 5 -numeroiset luvut tulkitaan postinumeroiksi). Tämän perusteella ensimmäisen tutkimuksen voidaan ajatella edustavan paikkatiedon esiintymismäärien alarajaa ja jälkimmäisen ylärajaa. Pienimmilläänkin paikkatietoa on kuitenkin sen verran, että sitä hyödyntäviä sovelluksia on mielekästä suunnitella.

5 Sijaintitiedon hyödyntäminen hakukoneissa

Hakusanoihin perustuvien hakukoneiden käyttö on eräs suosituimmista www:n hyötykäytön muodoista. Yhdestä tai useammasta hakusanasta muodostetaan hakuehto, jolle haetaan www:n sisällöstä mahdollisimman täydellistä vastaavuutta. Sijaintitiedon käyttö hakuehtona on sen sijaan varsin vähän käsitelty aihe [10]. Tavalliset hakukoneet eivät suoraan tue tämänkaltaista käyttöä. Käyttäjän sijaintitietoja hyödyntävä hakukone olisi erityisen hyödyllinen liikkuvalla käyttäjälle. Käyttäjä voisi esimerkiksi matkustaessaan hakea sijaintiaan lähellä olevia kohteita ja palveluja.

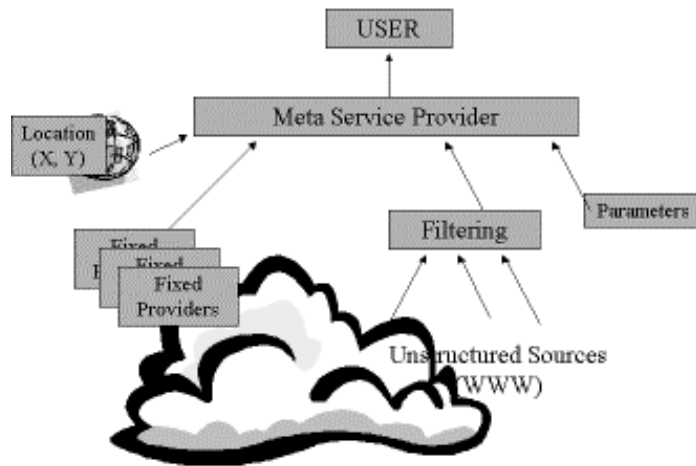
Hariharan, Fränti ja Mehta esittelevät artikkelissaan [10] idean mobiilista päätelaite-sovelluksesta ja palvelusta, joka tarjoaa tietoa käyttäjän sijainnin läheisyydessä olevista kiinnostuksen kohteista. Artikkelissa on visioitu tilanne, jossa käyttäjä voi lyhyillä komennoilla saada näkyviinsä tietoa haluamistaan kohteista sekä tilata tai varata palveluita. Sovelluksen mahdollistaa palvelu, joka käyttää hyväkseen käyttäjän sijaintia sekä tietoa hänen mieltymyksistään. Tietoa haetaan sekä vapaasta www:stä että useilta palvelusivustoilta. Löydetty tieto yhdistellään ja muotoillaan mobiililaitteella helposti käytettäväksi. Sovellus pyrkii myös oppimaan käyttäjänsä hakutoiveista, sekä automatisoimaan toimintaansa mahdollisimman paljon.

Artikkelissa sijaintirelevantti tiedonhakuprosessi on jaettu seuraaviin vaiheisiin:

1. Käyttäjän sijainnin määrittäminen.
2. Hyödyllisten tietolähteiden löytäminen. Näitä voivat olla tunnetut palvelutarjoajat (*fixed providers*) tai riippumattomat www-sivut.
3. Relevantin tiedon eristäminen tietolähteistä.
4. Tiedon prosessointi päätelaitetta varten.

Kuvassa 5 on hahmoteltu artikkelissa esitelty *Meta Service Provider*, joka kerää käyttäjän kiinnostuksen kohteisiin ja sijaintiin liittyvää tietoa erilaisista tietovarastoista.

Artikkelissa kuvattu sovellus on laaja ja koko maapallon alueella toimiva. Tässä tutkielmassa on tutkittu mahdollisuuksia toteuttaa suppeasti joitakin artikkelin ideoita käytännössä. Käytännön toteutuksen helpottamiseksi sovelluksen toiminta-alueena on



Kuva 5: Sijaintitietoa hyödyntävä hakupalvelu, Meta Service Provider [10].

vain Suomi. Samoin tietolähteinä käytetään vain suomalaisia www-sivuja. Seuraavissa aliluvuissa käydään läpi edellä olevan listan toimintojen toteutusmahdollisuuksia.

5.1 Käyttäjän sijainnin määrittäminen

Käyttäjän sijainti voidaan määrittellä monilla eri tavoilla (Luku 3.2). GPS-paikannus tarjoaa helposti toteutettavan menetelmän käyttäjän mobiilille ja reaaliaikaiselle paikannukselle. GPS-laitteesta saatavien koordinaattien käyttö hakukoneissa sellaisenaan ei ole kannattavaa, sillä www-sivuilla ei juurikaan esiinny koordinaatteja joihin käyttäjän sijaintia voisi verrata. Sensijaan osoitteiden ja paikannimien käyttö on yleistä (Luku 4). Tämän vuoksi käyttäjän sijainniksi kannattaa määrittellä koordinaattipistettä laajempi alue [10]. Käyttäjän tarkoille sijaintikoordinaateille on myöhemmin käyttöä löydetyn kohteen ja käyttäjän välisen etäisyyden määrittämisessä.

5.2 Tietolähteet

Sijaintiin perustuva palveluhaku on jo toteutettu useiden kaupallisten palveluntarjoajien (esimerkiksi Keltaiset sivut ja Fonecta) toimesta. Näissä hakutuloksissa näkyminen tosin edellyttää yhteisöltä tai yritykseltä maksua palveluntarjoajalle. Palvelut toimivat hyvin varsinkin silloin, kun haetaan tyypillisiä kaupallisia palveluita tai kohteita

(esimerkiksi ravintolat). Käyttäjän kiinnostus ei kuitenkaan välttämättä kohdistu pelkästään kaupallisiin kohteisiin, vaan hän saattaa kaivata tietoa vaikkapa jonkin aatteellisen yhdistyksen lähellä sijaitsevista toimipaikoista. Tällöin apu saattaa löytyä tavallisten www-sivujen joukosta.

5.2.1 Tiedonhaku www-sivuilta

Halutun kohteen löytäminen kaikkien maailman www-sivujen joukosta on valtaisa tehtävä. Ensiksi käytössä on oltava tehokas ja luotettava hakukone. Toiseksi haku tulee rajata niin, että kohteen löytämiseksi tarvitsee tutkia vain kohtuullinen määrä potentiaalisia sivuehdokkaita. Tässä tutkielmassa toteutetussa koesovelluksessa käytetään www-tiedonhakuun Google hakukonetta.

Google hakukoneen kehittäjiä olivat Stanfordin yliopiston opiskelijat Larry Page ja Sergey Brin [7]. Hakukoneen laskentateho perustuu suureen joukkoon toisiinsa linkitettyjä PC-tietokoneita. Toinen Googlen erikoisuus on *PageRank*-tekniikka. Tekniikan tehtävänä on määritellä hakuehtoja vastaavan sivun merkittävyys sitä suuremmaksi, mitä enemmän siihen on viitattu toisilta sivuilta. Myös viittaavan sivun merkittävyys vaikuttaa lopputulokseen. *PageRank*:in lisäksi käytetään *hypertekstin vastaavuusanalyysiä*, jossa sivuehdokkaan koko teksti käydään läpi. Näin tutkitaan sitä kuinka usein hakuehtoja vastaavat sanat esiintyvät ja missä kohdassa sivun tekstiä ne sijaitsevat.

Google ei tue sijaintitiedon käyttöä tavallisen www-haun hakuehtona. Koesovellus pyrkii hyödyntämään käyttäjän sijaintia siten, että tarkat sijaintikoordinaatit muutetaan sijaintia vastaavan lähimmän kunnan nimeksi. Tätä nimeä käytetään hakua rajaavana ehtona. Kun käyttäjä haluaa hakea esimerkiksi lähellä olevia ravintoloita, hän syöttää sovellukselle hakusanan ”Ravintola”. Sovellus muodostaa käyttäjän sijainnista ja hakusanasta hakuehdon, joka etsii ”Ravintola”-hakusanan ja käyttäjän sijaintikunnan nimen sisältäviä www-sivuja.

Hakujen rajaamisella sijaintikunnan avulla on kaksi päätavoitetta. Ensiksi on jotenkin kyettävä rajoittamaan saatavien tulossivujen määrää. Toiseksi, koska etsittävän kohteen ja käyttäjän välinen etäisyys pitää pysyä määrittämään, tulee tulossivun sisältää osoite. Sivuilla joilla esiintyy sijaintikunnan nimi, löytyy myös suuremmalla todennäköisyydellä osoitteita.

5.2.2 Tiedonhaku tunnetuilta palveluntarjoajilta

Erilaiset palveluntarjoajat tarjoavat hakupalveluita, joiden avulla on mahdollista löytää tietyllä alueella sijaitsevia, hakusanaa vastaavia kohteita. Haut suoritetaan kiinteästä tietokannasta ja tulokset saadaan nopeasti. Tarjolla on usein myös tarpeellista lisätietoa, kuten liikkeiden aukioloaikoja, karttoja ja reittiohjausta. Palvelut toimivat hyvin tyypillisten kaupallisten palveluiden hakemisessa. Yritysten ja yhteisöiden on kuitenkin maksettava näkymisestään palvelun tietokannassa. Tämän takia palveluhaut sopivat huonosti epäkaupallisten kohteiden hakemiseen. Palvelut eivät myöskään kykene huomioimaan käyttäjän tarkkaa sijaintia kohteiden esittämisessä, vaan johonkin paikkakuntaan liittyvät tulokset esitetään aina samassa järjestyksessä.

Koesovelluksessa on käytetty yhden kaupallisen palveluntarjoajan, Fonectan hakupalvelua. Hakuehto muodostetaan samalla tavalla kuin Googlen [www-hakujen](#) yhteydessä.

5.3 Relevantin tiedon eristäminen tietolähteistä

Hakukoneiden tuottamat tulokset perustuvat hakusanojen osalta merkkijonojen vastaavuuksien havaitsemiseen. Vaikka hakukoneen rajaavana ehtona käytetään kunnan nimeä, ei hakukone pysty tunnistamaan rajausehdon maantieteellistä luonnetta. Jos koesovelluksen hakuehdossa esiintyy sana ”Joensuu”, hakukone todennäköisesti palauttaa sivuja, jotka sisältävät tietoa suomalaisesta kaupungista. Todennäköistä myös on, että tulossivujen joukossa on kirjailija Matti Yrjänä Joensuuhun liittyviä sivuja. Pelkän kunnan nimellä ei tulossivuilla olevan tiedon relevanttiutta kyetä arvioimaan. Relevantin tiedon ohjelmallinen eristäminen vapailta [www-sivuilta](#) onkin hankalaa.

Koska [www-sivuilta](#) mahdollisesti löytyvien kohteiden etäisyys käyttäjästä mitataan, on kohdesivulta löydyttävä osoitteita. Osoite voidaan muuntaa koordinaateiksi sopivan tietokannan avulla, ja siten etäisyyden laskeminen mahdollistuu. Tämän perusteella määritellään relevanteiksi sivuiksi vain osoitteita sisältävät sivut.

Koska sovellusta käytetään mobiililla päätelaitteella, tulee huomioida myös tiettyjä rajoituksia: laitteiden käyttämät näytöt ovat pieniä ja tiedonsiirtonopeus usein melko vaatimaton [10]. Myös muut resurssit, kuten muistin määrä ja prosessorin teho, jäävät tavallisia PC-laitteita pienemmiksi. Tämän takia sovelluksen tulisi kyetä näyttämään

vain olennainen tieto ilman vaivalloista kokonaisten www-sivujen selailua. Niinpä sovelluksen pitää sivun relevanttiuden lisäksi kyetä määrittelemään ja eristämään relevantti tieto myös tämän sivun sisältä.

Koesovelluksen tapauksessa jokaisen hakukoneen tuottaman tulossivun lähdekoodi käydään läpi sana sanalta. Ensiksi pyritään löytämään osoitteisiin viittaavia sanoja. Kun tällainen sana löytyy, pyritään muodostamaan kokonainen osoite. Pelkkä osoite ei kuitenkaan tarjoa vielä riittävästi tietoa, joten sovellus pyrkii etsimään myös tähän osoitteeseen liittyviä kuvailutietoja. Osoite ja kuvailutieto muodostavat yhdessä hakutuloksen.

Kaupallisen palveluntarjoajan tapauksessa relevantin tiedon määrittäminen on helpompaa: onnistuneen haun kaikki tulokset ovat relevantteja. Tieto haetaan tietokannasta ja esitetään vakiomuotoisena, joten tarvittavan ja relevantin tiedon eristämistä ei ole vaikeaa automatisoida.

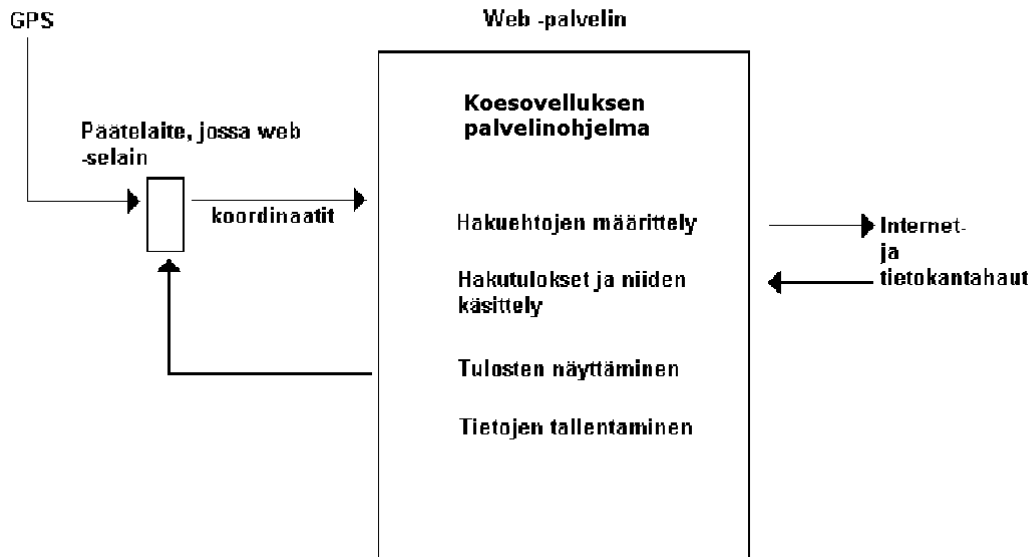
5.4 Tiedon prosessointi päätelaitteelle

Mobiilin päätelaitteen käyttö asettaa vaatimuksia esitettävän tiedon määrälle ja laadulle. Tiedon tulisi olla relevanttia, helppokäyttöistä ja nopeasti saatavilla. Tämän takia pelkkien relevanttien www-linkkien tuottaminen käyttäjälle ei ole riittävä ratkaisu. Lisäksi tulokset on myös järjestettävä nousevasti käyttäjän ja kohteen välisen etäisyyden mukaan.

Koesovelluksen tapauksessa löydettyistä hakutuloksista muodostetaan järjestetty lista. Jokainen listan alkio pyrkii sisältämään kohteen osoitteen, osoitteeseen liittyvän kuvausosion sekä kohteen ja käyttäjän välisen etäisyyden. Www-sivuilta saaduissa tiedoissa on lisäksi linkki sivulle, josta tieto on löydetty. Kaupallisen palveluntarjoajan tapauksessa mukana on linkki kohteeseen liittyviin lisätietoihin tai -palveluihin.

6 Koesovelluksen toteutus käytännössä

Koesovellus on ohjelma, joka hakee käyttäjän antamien hakuehtojen mukaista kohdetta sekä järjestää löydetty kohteet nousevasti etäisyyden perusteella. Hakusanojen lisäksi sovellus hyödyntää käyttäjän sijaintia yhtenä hakukriteerinä. Sovelluksen suunnittelun lähtökohtana ovat luvussa 5 esitellyt ideat käyttäjän sijaintitietoja hyödyntävästä palvelusta.



Kuva 6: Koesovelluksen rakenne.

Kuvassa 6 on esitetty koesovelluksen rakenne käyttäjän kannalta nähtynä. Sovelluksen käyttöliittymänä toimii www-sivu, joten ohjelman käyttöä varten tarvitaan www-selaimella varustettu päätelaite. Päätelaitteessa oleva sovelluksen osa noutaa käynnistyessään GPS-laitteesta koordinaatit, jotka sitten lähetetään palvelimelle. Ohjelman pääsivulla käyttäjä voi valita joko hakusanoihin tai aiemmin tallennettuun hakuprofiiliin perustuvan haun. Hakutulosten valmistuttua ne näytetään käyttäjälle. Mikäli hakutulokset ovat onnistuneita, voi käyttäjä tallentaa ne tietokantaan. Samalla tarjotaan mahdollisuutta tallentaa hakuehdot johonkin aiempaan tai uuteen hakuprofiiliin.

Kaikki ohjelman keskeisimmät toiminnot suoritetaan www-palvelimella. Tällä tavoitellaan mahdollisimman suurta laiteriippumattomuutta. Toisaalta tämänkaltainen toteutus vaatii päätelaitteelta yleisimpiä standardeja tukevaa www-selainta. Nykyaikaisista

internet-yhteyksillä varustetuista päätelaitteista se onneksi lähes aina löytyy. Www-toteutuksella voidaan myös pienentää prosessointitehoon, muistin kokoon ynnä muihin päätelaitteen resursseihin kohdistuvia vaatimuksia. Lisäksi ohjelman tekemät lukuisat haut internetistä sujuvat nopeammin palvelimelta palvelimelle kuin palvelimien ja mobiilin päätelaitteen välillä.

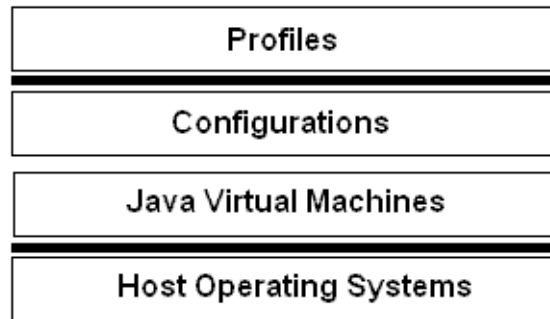
6.1 Ohjelmointitekniikat ja laitteisto

Päätelaitteessa sijaitsevan sovelluksen osan tehtävänä on luoda yhteys GPS-laitteeseen sekä lukea laitteen lähettämästä tiedosta tämänhetkiset sijaintikoordinaatit. Tämän jälkeen sovellus lähettää koordinaatit palvelimelle, avaten samalla päätelaitteen www-selaimen varsinaisen sovelluksen käyttöä varten.

Päätelaitesovelluksen ohjelmointikieleksi on valittu *J2ME* (Java 2 Micro Edition). Se on Java-ohjelmointikieleen perustuva ohjelmointiympäristö, joka on erityisesti suunniteltu rajoitetuilla resursseilla varustettujen laitteiden ohjelmointiin [13]. Koska tällaiset laitteet saattavat poiketa toisistaan suuresti (esim. matkapuhelin ja pesukone), sisältää J2ME erilaisia kokoonpanoja (configurations) erilaisille laitteistoille. Fyysisesti samankaltaiset laitteistot voivat kuitenkin poiketa täysin käyttötarkoituksiltaan. Näiden erojen huomioimiseksi J2ME:ssä käytetään profileja (profiles). Profilien ja kokoonpanojen alapuolella toimii JVM (Java Virtual Machine). JVM toimii Java-ohjelmien ja kulloisenkin laitteen käyttöjärjestelmän välissä. JVM on avain Java-ohjelmien laiteriippumattomuuteen. Kuvassa 7 on graafinen esitys J2ME:n arkkitehtuurista.

Ensimmäinen kokonaan toteutettu J2ME-profiili on *MIDP* (Mobile Information Device Profile). Profiili on suunniteltu erityisesti älypuhelin- ja PDA-käyttöön, joten sitä on käytetty päätelaitesovelluksen ohjelmoinnissa. Käytetty J2ME kokoonpano on puolestaan *CLDC* (Connected Limited Device Configuration), joka sisältää internetyhteydellä varustettujen ja pieniresurssisten laitteiden ohjelmointiin soveltuvia kirjastoja.

Koesovellus tärkein osa sijaitsee www-palvelimella ja se on ohjelmoitu *PHP*-kielellä, jonka loi Rasmus Lerdorf vuonna 1995 [19]. Alunperin kysymyksessä oli vain joukko Perl-skriptejä, joilla tekijä helpotti kotisivujensa ylläpitoa. Myöhemmin hän loi paljon laajemman C-kielisen toteutuksen, jota alkoi sitten julkaista avoimen lähdekoodin periaatteella. Tätä kirjoitettaessa PHP-kieli on edennyt versioon 5 ja sen jatkokehityksestä vastaa aktiivinen kehittäjäryhmä. Kielellä on helppo luoda dynaamisia verkkosivuja



Kuva 7: J2ME tasot [13].

tietokantayhteyksineen. Kielen viimeisimmät versiot tukevat myös olio-ohjelmointia.

Koesovellusta on tässä tutkielmassa käytetty laitteistolla, joka koostuu Nokia 6630 puhelimesta ja taskukokoisesta Pretec Bluetooth GPS Mini -laitteesta. Puhelimessa on valmiudet www-selailuun, Java-tuki sekä mahdollisuus ottaa yhteyksiä muihin laitteisiin *Bluetooth*-yhteyden avulla. Bluetooth on standardoitu, radioaaltoja käyttävä lyhyen matkan yhteystekniikka. Bluetooth-yhteyttä käytetään tässä sovelluksessa puhelimen ja GPS-laitteen väliseen kommunikointiin. Puhelimen Java tuki mahdollistaa omien Java-ohjelmien lataamisen puhelimen muistiin. Laitteiston fyysinen koko mitataavalla havainnollistettuna kuvassa 8.



Kuva 8: Koesovelluksen testeissä käytetty laitteisto.

6.2 Hakuprofiili

Hakuprofiilin tarkoituksena on helpottaa sovelluksen käyttöä. Sovelluksessa on mahdollista käyttää myös erillisiä hakusanoja, mutta ainakin testikokoonpanossa on kätevinä käyttää valmista profilia. Tällöin välttyy hankalalta ja virhealttiilta kirjoitustyöltä ohjelmaa käytettäessä. Profiili on yksinkertaisesti vain tallennettu lista hakusanoja, jotka käyttäjä voi nimetä kuvaavasti. Profilia käytettäessä ohjelma suorittaa peräkkäiset haut jokaisella profiiliin kuuluvalla hakusanalla. Profileita voi olla useita erilaisia.

Kun käyttäjä käyttää ohjelmaa ensimmäisen kerran, hänellä on mahdollisuus ainoastaan hakusanojen käyttöön. Kun hakutulokset ovat näkyvillä, käyttäjä voi arvioida haun onnistumista. Jos hakutulokset olivat hyödyllisiä, hän voi halutessaan tallentaa ne tietokantaan. Samalla tarjoutuu mahdollisuus nimetä uusi profiili, johon hakusana liitetään.

Myöhemmillä käyttökerroilla käyttäjä voi valita profiilin tai hakea edelleen hakusanoilla. Jos hakutulokset ovat onnistuneita, voidaan hakusanat tallentaa joko johonkin jo olemassaolevaan, tai uuteen profiiliin.

Profileja käytettäessä hakutulokset tallennetaan aina automaattisesti tietokantaan. Tämä mahdollistaa nopean uusintahaun esimerkiksi saman kaupungin toisella laidalla. Jos alue on suuri ja haluttuja kohteita eri osissa kaupunkia, voidaan lähimmän kohteen etsintä uuden sijainnin suhteen suorittaa nopeasti.

6.3 Kohteiden haku verkosta

HTTP (Hypertext Transfer Protocol) on internet protokolla, joka mahdollistaa www-sivujen lataamisen palvelimelta oman selaimen käytettäväksi [1]. Protokollan avulla on mahdollista myös välittää tietoa palvelimelta toiselle. Koesovelluksen www-palvelimella olevat osat hyödyntävät protokollan *GET*- ja *POST*-metodeja tietoja kerätessään [28]. *HTTP*:n *GET*-metodia käyttämällä voidaan ladata palvelimelta haluttu *www*-sivu tai sen tiedot. Metodi mahdollistaa myös palvelinosoitteen perään liitettävän pienen tieto-osan, joka käsitellään palvelimella, minkä jälkeen tulokset palautetaan *www*-dokumenttina. Merkkijonoa, jossa palvelinosoitteen perään liitetään tietoa, kutsutaan *kyselymerkkijonoksi*.

6.3.1 Www-haku Googlella

Koska koesovellus hakee tietoa myös vapaasta www-ympäristöstä, tarvittiin tehokas hakukone. Yksinkertaisin ratkaisu oli käyttää sovelluksen sisällä Google hakukonetta.

Kun käyttäjän koordinaatit on saatu syötteenä palvelimelle, niitä verrataan tietokannassa olevien suomalaisten kuntakeskusten koordinaatteihin. Lähimpänä käyttäjän sijaintikoordinaatteja oleva kunta valitaan hakuehdon osaksi. Tämä hakuehdon osa lähetetään Googlen palvelimelle. Sitä käytetään rajaamaan tulokset niihin suomalaisiin www-sivuihin, joissa se esiintyy täsmällisesti samassa muodossa. Sivuilta haetaan paitsi kohdetta tai palvelua, myös tarkkaa sijaintitietoa. Sijaintitieto on www-dokumenteissa tyypillisesti osoitteen muodossa (vrt. luku 4), joten täsmällinen kunnan nimi todennäköisesti esiintyy haluttua tietoa sisältävillä sivuilla.

Kuten aiemmin todettiin, käyttäjä voi antaa sovellukselle syötteenä joko hakusanan tai profiilin. Käyttäjän sijainti määritellään samaan aikaan kunnan tarkkuudella. Hakusana (joko profiilista tai käyttäjän määrittämänä) sekä kunta muodostavat yhdessä hakuehdon, jota käytetään Google hakukoneen hakusanoina. Haku käynnistetään lähettämällä Googlen palvelimelle kyselymerkkijono. Merkkijono käynnistää Google-haun seuraavilla parametreille:

- Ensimmäinen hakusana: Voi muodostua useista sanoista, jolloin haetaan kaikilla sanoilla (+ operaattori).
- Toinen hakusana: kunnan nimi.
- Haetaan suomenkielisiltä sivuilta.
- Haetaan maksimissaan 10 linkkiä.

Ensimmäinen hakusana voi siis koostua myös useista sanoista (esim. kiinalainen ravintola). Haku hakee sivuja, jossa esiintyvät kaikki ehdossa olevat sanat. Hakutuloksiin tulee myös sivuja, joissa sanat ovat eri järjestyksessä tai esiintyvät toisistaan erillisinä. Toisen hakusanan tulee esiintyä sivulla tarkalleen samassa muodossa. Ehdon tarkoituksena on rajata hakua sellaisiin sivuihin, joista sijaintitiedon löytyminen on todennäköistä.

Haku rajataan koskemaan vain suomenkielisiä www-sivuja. Tämä johtuu siitä, että sijaintitiedon löytämiseen käytetyt menetelmät perustuvat suomalaiseen osoite- ja pos-

tinumerojärjestelmään. Lisäksi sijaintien vertailu etäisyyden perusteella tapahtuu suomalaisessa kkj -yhtenäiskoordinaatistossa.

Viimeinen rajausta koskee Googlen palauttamien hakutulosten määrää, joka on rajoitettu enintään kymmeneen www-sivuun. Koska jokaisen linkin osoittama sivu ladataan palvelimelle ja sen koko lähdekoodi käydään ohjelmassa läpi, ohjelman toiminnan nopeus riippuu tutkittavien sivujen määrästä. Jos näyttää siltä, että olennaisia sivuja jää pois, voidaan parametriä muuttaa. Google sallii maksimissaan sadan linkin tuloslistan.

Seuraavassa on ohjelman Google-palvelimelle lähettämä kyselymerkkijono. Muuttujassa "num" on haettavien sivujen maksimimäärä, "\$txtServices" sisältää hakusanat ja \$txtPlace kunnan nimen.

```
http://www.google.fi/search?as_q=$txtServices&num=10&hl=fi
&btnG=Google-haku&as_epq=$txtPlace&as_oq=&as_eq=&lr=lang_fi
&as_ft=i&as_filetype=&as_qdr=all&as_occt=any&as_dt=i
&as_sitesearch=
```

Kun edellä kuvatulla tavalla parametroitu kyselymerkkijono lähetetään Googlen palvelimelle, se palauttaa hakua vastaaville sivuille ohjaavat linkit. Linkit tallennetaan taulukkoon. Taulukko käydään läpi lataamalla jokaisen linkin edustama sivu ohjelman tutkittavaksi. Tässä vaiheessa tehdään vielä muutamia rajaamistoimenpiteitä. Sivujen suurin mahdollinen koko on rajoitettu sekä datamäärän että rivimäärän suhteen. Samoin tutkittaviksi dokumenteiksi hyväksytään vain html-dokumentit. Ohjelma tallentaa kunkin sivun lähdekoodin sanataulukoksi (PHP-array), jonka koko ei voi kasvaa rajattomasti. Tutkittavaksi joutunut pdf- tai doc-päätteinen dokumentti saattaa sisältää kymmeniä tekstisivuja, jolloin sanataulukon kokoraja voi ylittyä.

6.3.2 Haku Fonectan tietokannasta

Fonecta Oy on suomalainen yhteystietoyritys, joka tarjoaa puhelinluetteloita ja puhelin-, matkapuhelin- sekä internetpalveluita [5]. Internetpalvelut sisältävät sekä erilaisia maksullisia yhteystietohakua että ilmaisen www.020202.fi -palvelun. Hakusivu tarjoaa mahdollisuuden etsiä palveluita paikkakunnan nimellä, hakea reittejä ohjeineen ja karttoineen sekä erillisen ravintola- ja sähköpostihaun. Koesovellus käyttää omassa tiedonhaussaan paikkakuntakohtaista palvelu- ja yrityshakua.

Hakuehdot muodostuvat käyttäjän antamana samalla tavalla kuin Google-haun tapauksessa. Kyselymerkkijonon sijasta hakuehdot lähetetään HTTP-protokollan *POST*-metodilla [28]. Tässä tekniikassa palvelimelle lähetetään tiedot hakusanoista ja pyydetään palvelinta käsittelemään ne. Itse asiassa sovellus näyttäytyy tällöin www.020202.fi palvelimelle samalla tavalla kuin sen oma asiakassivu.

Jos haku onnistuu, palauttaa palvelin vakiomuotoisen vastaussivun, joka sisältää oletuksena vähintään 10 hakutulosta. Jos hakutuloksia on enemmän, ensimmäisellä sivulla on niitä sisältäville sivuille linkit. Hakutulokset sisältävät kaikki olennaiset yhteystiedot, sekä valinnaisen kuvausosion, jossa voi laajemmin kuvailla kohteen ominaisuuksia. Lisäksi jokaiseen hakutulokseen liitetään medialinkkejä, joiden avulla voi saada selville esimerkiksi kohteeseen liittyvät mahdolliset *www*-sivut, aukioloajat tai nähdä sen sijainnin kartalla.

6.4 Paikkatiedon eristäminen

Eräs koesovelluksen perustavoitteista oli kyetä järjestämään käyttäjän hakujen mukaiset palvelut ja kohteet etäisyysjärjestykseen. Jotta tämä olisi mahdollista, pitää *www*-sivuilta onnistua eristämään sijaintitietoa. Käytännössä tämä tarkoittaa täydellisten osoitteiden löytämistä. Luvussa 4 esitelty geo- ja address-tagien käyttö on vielä siinä määrin harvinaista, että tarkka sijainnin määrittely perustuu osoitteisiin.

Sijainti olisi mahdollista määritellä myös kunnan tai postinumeroalueen tarkkuudella. Tällöin kuitenkin menetetään mahdollisuus järjestää etäisyyden perusteella useampia, samalla paikkakunnalla sijaitsevia kohteita. Riittävän tarkka sijainti saadaan selville, jos tekstistä kyetään löytämään katuosoite. Katuosoitteen ja sijaintikunnan perusteella voidaan hakea sijaintikoordinaatit osoite/koordinaatti -tietokannoista. Tällaisia tietokantoja löytyy Suomesta useita. Koesovellus käyttää tähän tarkoitukseen maanmittauslaitoksen ”Kansalaisen karttapaikka” *www*-sivustoa.

Pro-gradu tutkielmassaan Ilkka Vänskä tutki sovelluksellaan erilaisten paikkatietoelementtien esiintymistiheyttä *www*-dokumenteissa. Koesovellus käyttää siinä esiteltyjä ideoita osoitteiden tunnistamiseen muun tekstin joukosta. Vänskän sovelluksessa *www*-sivun teksti pilkottiin ensin sanataulukoksi. Taulukon sanat käytiin sen jälkeen läpi yksi kerrallaan. Sanoja tutkimalla pyrittiin havaitsemaan oliko kysymyksessä esimerkiksi katuosoite, kunnan- tai paikannimi, postinumero tai puhelinnumero.

Koesovellus eroaa edellisestä siten, että yksittäisen paikkatietoelementin sijasta pyritään löytämään kokonaisia osoitteita. Toisin sanoen, on löydettävä useita toisiinsa liittyviä, lähekkäin sijaitsevia paikkatietoelementtejä. Paikkatietoon kuuluu sijaintitiedon (siis osoitteen) lisäksi myös muuta tietoa. Jotta löydetyllä osoitteella olisi käyttäjälle merkitystä, pitäisi löytää tekstiä kuvailemaan myös sitä asiayhteyttä, johon osoite liittyy.

Www-sivun pilkkominen sanataulukoksi aloitetaan poistamalla sivun lähdekoodista mahdollisimman tarkoin kaikki ylimääräinen. Www-sivut sisältävät paljon tekstin muotoiluun ja sivun toimintoihin liittyviä määrittelyksiä. Näihin liittyviä merkkijonoja ei ole tarpeen tutkia koesovelluksessa. Testausvaiheessa havaittiin, että lähdekoodin ”siivouksella” tutkittavan tekstin määrä vähenee lähes 40%. Tästä huolimatta tutkittavaan tekstiin jää vielä paljon ylimääräisiä merkkijonoja.

Tekstin käsittelyssä joudutaan noudattamaan varovaisuutta, jotta poistettavien sanojen mukana ei menetettäisi myös tarpeellista tietoa. Sovelluksen ohjelmoinnissa käytetty PHP-kieli sisältää kyllä funktion muun muassa tagien poistamiseen tekstistä, mutta sen käyttäminen aiheutti myös olennaisen tiedon menetyksiä. Syynä tähän oli yleensä lopetustagin puuttuminen tutkittavalta lähdekoodiriviltä. Koesovellus tutkii lähdekoodin rivi riviltä ja poistaa aloitus- ja lopetustagit sekä niiden välissä olevan tekstin. Tämä onnistuu vain, jos kaikki tagin osat löytyvät tutkittavalta riviltä. Poistetut osat muutetaan välilyönneiksi. ”Puhdistetut” rivit pilkotaan sanataulukoksi välilyöntien perusteella. Tässä vaiheessa on otettava huomioon myös skandinaavisten merkkien muuntaminen ja esimerkiksi väliviivalla yhdistettyjen sanojen määrittely yhdeksi sanaksi.

Paikkatiedon eristäminen keosovelluksessa alkaa katuosoitteeseen viittaavien sanojen etsinnällä sanataulukosta. Kun käsittelyvuorossa oleva sana loppuu merkkijonoon ”katu”, ”kuja”, ”polku”, ”raitti” tai ”tie”, yritetään muodostaa osoite. Seuraavan viiden sanan joukosta pyritään löytämään postinumero. Jos löytyy viisi numeroa sisältävä merkkijono, selvitetään postinumerotietokannan avulla onko kyseessä postinumero. Jos on, muodostetaan osoitteen loppuosa postitoimipaikan ja kunnan osalta tietokannasta löytyvistä tiedoista. Samassa yhteydessä tallennetaan katuosoitetta kuvaavan päätteiden sisältämän sanan ja postinumeron väliset merkkijonot katuosoitteeksi. Jos osoite on kirjoitettu tyypillisessä suomalaisessa muodossa, saadaan näin muodostettua täydellinen osoite lähiosoiteineen. Lähiosoite, kuten ”Länsikatu 15”, on edellytys tarkan sijainnin määrittelemiseksi.

Seuraavaksi ohjelma yrittää löytää tekstiä, joka kertoisi käyttäjälle mihin löydetty osoite liittyy. Yksinkertainen tapa olisi liittää osoitetietoon sen www-sivun otsikko mistä kyseinen osoite löytyi. Valitettavasti monilla, varsinkin kaupallisia palveluja koskevilla hakusanoilla, saadaan tulokseksi erilaisia yhteystietopalveluiden yhteystietolistoja. Tällöin sivun otsikko kertoo vain mikä palvelu tiedot on kerännyt.

Toinen tapa olisi tallentaa ne rivit tekstistä, joissa hakusana tai hakusanat esiintyvät. Ongelmana ovat jälleen sivut, jotka sisältävät useita osoitteita. Tällöin ei voida tietää, mihin osoitteeseen näiden rivien sisältämä tieto itse asiassa liittyy. Esimerkiksi jos hakusanamme ovat ”pizzeria” ja ”Joensuu”, saamme tutkittavaksemme todennäköisesti sivun joka sisältää joensuulaisia yrityksiä monenlaisilta toimialoilta. Tällaisessa tapauksessa pizzeria-sanan sisältävien rivien liittäminen jokaiseen sivulta löydettyyn joensuulaiseen osoitteeseen antaisi käyttäjälle harhaanjohtavaa tietoa.

Kolmas tapa voisi olla sivun metatieto-osion tutkiminen. Luvussa 4 kerrottiin kuinka metatiedon avulla voidaan välittää dokumentin sisältöön liittyviä tietoja. Keräämällä tekstiä meta-elementtien ”description” ja ”keywords” -kentistä, voidaan käyttäjälle tuottaa kuvauksia dokumentin sisällöstä. Ongelmana ovat jälleen useiden alojen yhteystietoja tarjoavat sivustot. Löydetyt osoitteet eivät yksilöidy metatiedon avulla.

Koesovelluksen toteutusvaiheessa on lisätietojen etsinnässä testattu myös edellä mainittuja tekniikoita. Tulokset eivät vastanneet ohjelman toimivuudelle asetettuja vaatimuksia. Lopulta onnistuneimmaksi menetelmäksi osoittautui tekstistä löydettyä osoitetta edeltävien, lähimpien sanojen tallentaminen kuvailutiedoksi. Usein tarkan osoitteen edellä on siihen liittyvän kohteen nimi. Koesovellus tallentaa kuvailutiedoksi viisi katuosoitetta edeltävää sanaa.

Kun tekstistä on löydetty osoite-ehdokas, sille haetaan koordinaatit maanmittauslaitoksen www-sivuilta. Viimeistään tässä vaiheessa tulee tarkistetuksi onko kysymyksessä oikea käyttökelpoinen osoite. Jos osoitetta ei löydy tietokannasta, sille ei myöskään voida laskea etäisyyttä käyttäjästä. Tällaiset osoitteet on merkitty hakutuloksessa siten, että etäisyyden kohdalla on merkintä ”ei saatavilla”. Kun lopulliset hakutulokset tallennetaan tietokantaan, nämä osoitteet putoavat pois. Varsinaisen osoitteen ohella hakutulokseen tallennetaan edellä aiemmin mainitun kuvailutiedon lisäksi etäisyys käyttäjästä, sijaintikoordinaatit sekä se www-osoite, josta tiedot löytyivät.

Kun haku kohdistuu Fonectan palvelusivustolle, on paikkatiedon eristäminen helpom-

paa. Hakutuloksia palauttavat sivut ovat vakionmuotoisia, joten osoitteiden ja niihin liittyvien otsikkotietojen kerääminen onnistuu luotettavasti. Lisäksi Fonecta tarjoaa jokaiselle kohteelle erillisen medialinkin, jota seuraamalla löytyy lisätietoja kohteista. Tämä www-osoite tallennetaan hakutuloksien ”Linkki” -kohtaan. Linkin osoittama sivu kertoo esimerkiksi liikkeiden aukioloajat, sekä näyttää myös kohteen sijainnin kartalla.

6.5 Ohjelman käyttämät tietovarastot

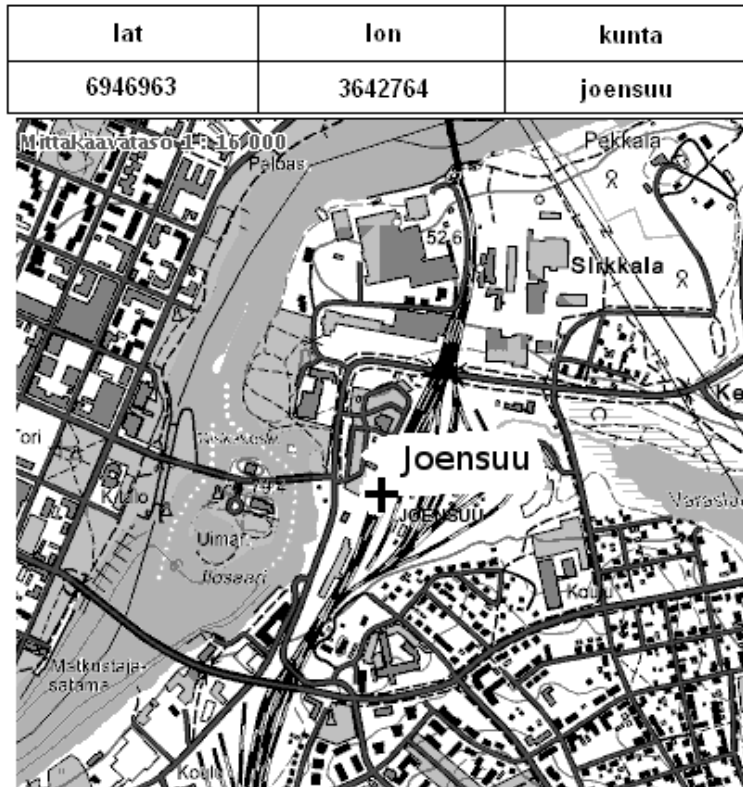
Koesovellus hyödyntää käyttäjän sijaintitietoja kahdella tavalla. Ensiksi GPS-koordinaattien perusteella määritellään lähimpänä käyttäjää oleva suomalainen kuntakeskus. Toiseksi koordinaatteja käytetään haun tuloksena löydettyjen kohteiden etäisyyden mukaista järjestämistä varten.

Kun käyttäjä aloittaa koesovelluksen käytön www-palvelimella, hänen sijaintikoordinaattinsa tutkitaan. Ohjelma suorittaa haun tietokantaan, joka sisältää kaikkien Suomen kuntien kuntakeskusten koordinaatit.

Kuntien koordinaatit on kerätty ”Kansalaisen karttapaikka” -palvelun avulla [14]. Palvelun toiminto ”haku paikannimellä” tuottaa kunnan nimeä käytettäessä yleensä vähintään kaksi karttapistettä. Usein ensimmäisenä on kyseisen kunnan rautatieasema ja toisena kuntakeskuksen maantieteellinen keskus. Koesovelluksen kuntakoordinaattitietokantaan on tallennettu listan ensimmäisenä oleva sijaintitieto. Tietokaannon kokoamiseen on käytetty kaikki Suomen kunnat sisältävää tiedostoa. Pienen PHP-ohjelman avulla on haettu jokaista kuntaa vastaavat koordinaatit ohjelmallisesti edellä mainitusta palvelusta. Tietokanta sisältää siis vain kunnan nimen, sekä sitä vastaavat kkj-yhtenäiskoordinaatiston mukaiset leveys- ja pituuskoordinaatit,

Kuvassa 9 on kuntien koordinaatteja sisältävästä tietokannasta Joensuun koordinaatit sisältävä rivi. Koordinaatteja vastaavasta karttapisteestä nähdään, että kyseessä on Joensuun rautatieasema.

Koesovellus käyttää Karttapaikan www-sivustoa tietovarastona myös dynaamisesti. Kun ohjelma löytää käyttäjän hakua vastaavia kohteita, niiden sijaintikoordinaatit haetaan ”haku osoitteella” -toiminnolla. Ohjelma lähettää Karttapaikan palvelimelle kohteen katuosoitteen ja kunnan nimen. Palvelu palauttaa maanmittauslaitoksen tietokannasta koordinaatit jos osoitteelle sellaisia löytyy. Koordinaatteja käytetään ohjelmassa



Kuva 9: Joensuuta edustavat koordinaatit ja niitä kuvaava karttapiste.

kohteen ja käyttäjän välisen etäisyyden laskemiseen.

Kun käyttäjä antaa koesovellukselle hakukomennon, sovellus tutkii ensin onko samoilla hakuehdoilla jo aiempia tuloksia. Hakukomennon hakusanoja sekä käyttäjän tämänhetkistä sijaintia verrataan hakutuloksia sisältävän tietokannan vastaaviin. Jos hakusanat ja sijaintia vastaava kunta täsmäävät tietokannassa oleviin, palautetaan käyttäjälle aiemmin haetut tiedot. Kohteiden etäisyys käyttäjästä lasketaan kuitenkin nykyisen sijainnin mukaan. Näin löydetty kohteet voidaan näyttää etäisyysjärjestyksessä käyttäjälle. Tietokannan käytöstä on hyötyä silloin, kun käyttäjä liikkuu esimerkiksi saman kaupungin eri alueilla. Jos käyttäjän kiinnostuksen kohteet pysyvät samoina, voidaan haku uudistaa ja saada tulokset nopeasti, oikeassa etäisyysjärjestyksessä.

Kuvassa 10 on esitetty Hakutulos-tietokantataulun rakenne sekä erään hakutuloksen tiedot sisältävä rivi. ”Hakuehto”-kenttään tallennetaan hakusana tai hakusanat, ”paikka” puolestaan kuvaa tämänhetkistä sijaintia lähimpänä olevaa kuntaa. Osoite tallennetaan täydellisenä ”lahios”, ”pnro”, ”ptoimipaikka” ja ”kunta” -kenttiin. ”Kuvaus” sisältää kohteeseen liittyviä muita tietoja. ”Linkki” on edellä mainitut tiedot sisältäneen

www-sivun osoite. Osoitteelle haetut koordinaatit on tallennettu ”lat” ja ”lon” -kenttiin kkj-yhteinäiskoordinaatisto -muodossa.

Hakutulos	
hakuehto	Pizzeria
paikka	joensuu
lahios	Niskakatu 11
pnro	80100
ptoimipaikka	JOENSUU
kunta	joensuu
kuvaus	KOTIPIZZA : puh 013 223228
linkki	http://www.020202.fi/media/117614
lat	6947300
lon	3641952

Kuva 10: Hakutulos-tietokantataulu.

Hakusanojen sijasta koesovelluksen käyttäjällä on mahdollisuus suorittaa haku- ja myös valmiilla hakuprofililla (luku 6.2). Hakuprofiilit on tallennettu Profilit-tietokantatauluun. ”Tunnus”-kenttä sisältää käyttäjän profilille antaman nimen ja ”hakusanat” puolestaan kaikki profiliin liitetyt hakusanat. Kuvassa 11 on ”Ruokapaikat”-profiliin tallennetut tiedot.

Profiilit	
tunnus	Ruokapaikat
hakusanat	Ravintola;Pizzeria;Kiinalainen+Ravintola

Kuva 11: Profilit-tietokantataulu.

Kun koesovellus etsii www-sivuilta osoitteita, se hyödyntää ”Paikat”-tietokantaa. Tietokantaan on tallennettu kaikki suomalaiset postinumerot sekä niihin liittyvät postitoimipaikat ja kunnat. Tietokannan avulla voidaan tarkistaa postinumeron oikeellisuus ja täydentää yhdenmukaisella tavalla hakutuloksen paikkatietoa postitoimipaikan ja kunnan osalta. Tietokannan tiedot on kerätty ohjelmallisesti Suomen Postin postinumeroluettelo-sivustolta [23]. Kuvassa 12 ovat tiedot postinumerolle 80100.

Paikat	
pnro	80100
ptoimipaikka	JOENSUU
kunta	joensuu

Kuva 12: Paikat-tietokantataulu.

Kun käyttäjä suorittaa uuden palveluun tai muuhun kohteeseen kohdistuvan haun, sovellus käyttää tietovarastona suomalaisia www-sivuja ja Fonectan palveluhaku -sivuja. Näistä tietovarastoista haettu tieto näytetään käyttäjälle ja tarvittaessa tallennetaan ”Hakutulos”-tietokantaan.

7 Koesovelluksen käyttökokemuksia

Koesovelluksen käyttökokeiluja varten suunniteltiin kolme erityyppistä testitilannetta:

1. Matkailija vieraassa kaupungissa. Suoritetaan kaupunkia koskevia tiedonhakuja.
2. Matkailija liikkeellä maaseudulla. Suoritetaan tiedonhakua lähimmistä kohteista ja palveluista.
3. Hakuprofilien toiminnan testaus.

Ensimmäisen tyypin testitapauksissa tarkoituksena oli jäljitellä tilannetta, jossa käyttäjä haluaa löytää itselleen vieraassa kaupungissa joitakin tyypillisiä matkailijan tarvitsemia palveluja. Näiden lisäksi käyttäjä tiesi kaupungista etukäteen joitakin eikaupallisia kohteita. Näitä haettiin joko täsmällisillä tai kuvaavilla hakusanoilla. Jokaisessa kaupungissa haettiin tässä yhteydessä lähimmän ortodoksisen kirkon, yliopiston tai siihen liittyvän laitoksen, raviradan sekä jonkin aatteellisen yhdistyksen tietoja. Kaupunkitestejä suoritettiin Joensuussa, Kuopiossa ja Helsingissä. Joensuussa käyttäjän sijainti oli tiedepuiston edessä, Kuopiossa Minna Canthinkadun päässä sekä Helsingissä Pasilan rautatieasemalla.

Toisen tyypin testitapauksilla pyrittiin jäljittelemään tilannetta, jossa käyttäjä on matkalla taajama-alueiden ulkopuolella. Testeissä haettiin hakusanoilla jälleen tyypillisiä matkailijan kiinnostuksen kohteita. Lisäksi jokaisessa tapauksessa yritettiin selvittää lähiseudulla sijaitsevan yrityksen sekä lähimmän eläinlääkärin sijaintia. Testi suoritettiin kolmella eri paikkakunnalla. Käyttäjän sijainnit olivat Herajoen tie Enossa, Maaherrantie Keiteleellä sekä Nurmijärventie Lieksassa.

Kolmannen tyypin testitapauksissa keskityttiin hakuprofilien muodostamiseen ja käyttöön. Tavoitteena oli todeta, kärsiikö ohjelman vakaus profiliin kuuluvien useiden hakujen perättäisestä suorittamisesta. Testattiin myös sitä, kuinka paljon hitaammin ohjelma toimii profileja käytettäessä. Sen lisäksi testauksessa havainnoitiin, kuinka hakutulosten laatu vaihtelee, kun samoilla profileilla haetaan tietoja eri puolilla Suomea.

Pohjois-Karjalan alueella testaus suoritettiin luvussa 6.1 kuvatulla laitteistolla. Muissa tapauksissa ohjelmaa käytettiin ilman GPS-yhteyttä lisäämällä syötekoordinaatit käsin.

7.1 Testituloksia

Kaikkien kolmen testitapaustyyppin kohdalla pidettiin kirjaa hakutulosten löytymisestä sekä niiden oikeellisuudesta. Kahdessa ensimmäisessä testitapaustyyppissä mahdollisia ongelmatilanteita ennakoitiin luomalla sarakkeet eri tavoin epäonnistuneiden hakujen kirjaamiseksi.

Ensimmäisessä taulukon sarakkeessa ovat käytetyt hakusanat. Fonecta- ja WWW-sarakkeet kertovat kuinka monta hakutulosta (osoite ja siihen liittyvät tiedot) kyseisistä tietovarastoista on löydetty. Numeroidut sarakkeet kertovat kuinka monta kunkin tyyppisiä virheitä hakutuloksista löytyi. Seuraavassa käytetyt virhetyypit sarakenumeroiden mukaisesti:

1. *Tiedon eristämisvirhe Fonectan sivulta.* Tämä tarkoittaa tilannetta, jossa ohjelma on epäonnistunut palvelun palauttaman hakutuloksen lukemisessa.
2. *Fonectan palauttaman hakutuloksen etäisyyttä ei voitu laskea.* Toisinaan tietokanta ei sisällä täydellistä osoitetta, joten koordinaatteja ei voida hakea.
3. *Www-sivulta löydetty hakutulos täysin epärelevantti.* Tämä tapaus tarkoittaa tilannetta, jossa saatu hakutulos ei liity mitenkään haettuun palveluun tai kohteeseen.
4. *Www-sivulta löydetty hakutuloksen paikkatieto epärelevantti, mutta mukana olevan linkki vie sivulle, josta oikean tiedon voi löytää.* Virhe aiheutuu yleensä www-sivusta, joka sisältää useita erityyppisiin kohteisiin ja palveluihin liittyviä osoitteita.
5. *Www-sivulta löydettyssä hakutuloksessa ei ole kunnollista kuvausosiota (otsikkoa).* Tällöin on löydetty osoite, mutta ei kunnollista kuvausta siitä, mihin osoite liittyy.

Viimeisessä sarakkeessa on hakutulosten hakemiseen kulunut aika. Kaikki yhteen sijaintiin liittyvät haut on tehty peräkkäin. Kaikki haut olivat ”uusia hakuja”, sillä tuloksia ei ollut valmiina hakutulostietokannassa. Aika on mitattu palvelimella sijaitsevan ohjelman osan toiminnalle. Tähän on vielä lisättävä aika, jonka sovellus käyttää GPS-laitteen koordinaattien hakemiseen ja kytkeytymiseen palvelimelle. Testikokoonpanol-

la tuo aika oli noin 20 sekuntia. Alimmalla taulukon rivillä on ilmoitettu hakutulosten ja virheiden yhteismäärät sekä hakuajojen keskiarvot.

7.1.1 Matkailija vieraassa kaupungissa

Taulukoissa 1, 2 ja 3 ovat yhteenvedot kaupunkialueilla suoritettujen hakujen tuloksista.

Taulukko 1: Hakutulosten määrät ja virheet virhetyypeittäin. Kaupunkialueen haku Joensuun alueella.

<i>Hakusana</i>	<i>Fonecta</i>	<i>WWW</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Aika[sek.]</i>
Ravintola	9	4		2			1	40
Pizzeria	10	6					5	35
Hotelli	7	8						35
Ortodok. kirkko	0	3			2		1	40
Autokorjaamo	10	3					1	35
Tiedepuisto	2	0						40
Pop-muusikot ry	0	27				23	1	35
Ravirata	1	6		1		6		35
<i>Yhteensä</i>	39	57		3	2	29	9	ka. 37

Taulukko 2: Hakutulosten määrät ja virheet virhetyypeittäin. Kaupunkialueen haku Kuopion alueella.

<i>Hakusana</i>	<i>Fonecta</i>	<i>WWW</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Aika[sek.]</i>
Ravintola	9	50	1	2			47	45
Pizzeria	10	19		1	4			35
Hotelli	10	6	1				2	35
Ortodok. kirkko	1	4			1	1		40
Autokorjaamo	9	19		2	4			40
Tietojenk.tieteen laitos	0	0						20
Kuopion Riento	0	18			2	15		35
Ravirata	1	0		1				40
<i>Yhteensä</i>	40	116	2	6	11	16	49	ka. 36

Taulukko 3: Hakutulosten määrät ja virheet virhetyypeittäin. Kaupunkialueen haku Helsingin alueella.

<i>Hakusana</i>	<i>Fonecta</i>	<i>WWW</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Aika[sek.]</i>
Kiinalainen ravintola	9	12		2				35
Pizzeria	10	51					40	45
Hotelli	10	5	4					45
Ortodok. kirkko	2	1		1				35
Autokorjaamo	7	107		3	1		1	50
Tietojenk.tieteen laitos	0	0						40
Helsingin jujutsuklubi	0	2				1	1	40
Ravirata	1	1			1			40
<i>Yhteensä</i>	39	179	4	6	2	1	42	ka.41

7.1.2 Matkailija liikkeellä maaseudulla

Taulukoissa 4, 5 ja 6 ovat yhteenvedot testin tuloksista samoin kuin edellisen luvun testitapauksissa. Hakusanat vastaavat nyt paremmin taajama-alueen ulkopuolella suoritettavaa tiedonhakua.

Taulukko 4: Hakutulosten määrät ja virheet virhetyypeittäin. Käyttäjä Herajoentiellä, Enossa.

<i>Hakusana</i>	<i>Fonecta</i>	<i>WWW</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Aika[sek.]</i>
Huoltoasema	2	3			3			30
Hinauspalvelu	2	0						40
Majoitus	5	7				2		45
Enon Erä ja Urheilu	0	3				1		35
Eläinlääkäri	1	4				4		30
<i>Yhteensä</i>	10	17			3	7		ka. 36

Hakuja edellä mainituissa testitapauksissa oli yhteensä 39. Saatujen hakutulosten yhteismäärä oli 632, joista Fonectan palvelusta 153 ja www-sivuilta 479. Hakuaiakeskiarvoksi tuli 36,4 sekuntia. Taulukoissa 7 ja 8 on esitetty virheellisten tulosten kokonaismäärät.

Taulukko 5: Hakutulosten määrät ja virheet virhetyypeittäin. Käyttäjä Nurmijärventiellä, Lieksassa.

<i>Hakusana</i>	<i>Fonecta</i>	<i>WWW</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Aika[sek.]</i>
Huoltoasema	4	8		1	3	5		35
Hinauspalvelu	1	2			2			35
Majoitus	8	25	1	5	1	13		40
Erästely ky	1	11				8		35
Eläinlääkäri	1	39			39			30
<i>Yhteensä</i>	15	85	1	6	45	26		ka. 35

Taulukko 6: Hakutulosten määrät ja virheet virhetyypeittäin. Käyttäjä Maaherrantiellä, Keiteleellä.

<i>Hakusana</i>	<i>Fonecta</i>	<i>WWW</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Aika[sek.]</i>
Huoltoasema	2	1			1			30
Hinauspalvelu	4	0						35
Majoitus	2	1						35
Masi Oy	1	21				18		30
Eläinlääkäri	1	2		1		1		35
<i>Yhteensä</i>	10	25		1	1	19		ka. 33

Taulukko 7: Virheellisten tulosten määrä Fonecta-hauissa

<i>Virhetyypin numero</i>	<i>[kpl]</i>	<i>[%]</i>
tyyppi 1	7	4,6
tyyppi 2	22	14,4

Taulukko 8: Virheellisten tulosten määrä www-hauissa

<i>Virhetyypin numero</i>	<i>[kpl]</i>	<i>[%]</i>
tyyppi 3	64	13,4
tyyppi 4	98	20,6
tyyppi 5	100	20,9

Tulosten perusteella voidaan todeta, että täysin hyödyttömiä hakutuloksia (virhetyypit 1 ja 3) oli Fonectan tuloksista 4,6% ja www-sivujen tuloksista 13,4%. Fonectan kohdalla tulos oli pettymys, sillä vakiomuotoisilta tulossivuilta tulisi pystyä eristämään paikkatieto virheettömästi. Toisaalta www-sivuilla saatujen hyödyttömien tuloksien määrän vähäisyys oli positiivinen yllätys. Tosin näin pienellä aineistolla ei voi vetää kovin pitkälle meneviä johtopäätöksiä.

Mielenkiintoinen ja samalla ylivoimaisesti eniten hyödyttömiä tietoa antanut yksittäinen haku oli Nurmijärventiellä Lieksassa tehty koe ”eläinlääkäri”-hakusanalla. Tässä tapauksessa täysin epärelevantteja tuloksia saatiin 39 kappaletta. Asiaa tutkittaessa osoittautui, että haku löysi Lieksan kaupungin sivun, jossa oli kaikkien kaupunginvaltuuston jäsenien nimet osoitteineen. Joukossa oli yksi eläinlääkäri, mutta myös 38 täysin asiaan liittymätöntä henkilöä ja osoitetta. Näin yksi ainoa www-sivu aiheutti lähes puolet koko testin epäonnistuneista www-hauista.

Virhetyypin 4:n osuus hakutuloksista oli noin 20%. Näissä tapauksissa voi selailuun kykenevällä päätelaitteella löytää relevanttia tietoa hakutulokseen sisältyvää www-linkkiä seuraamalla. Myös virhetyypin 5 osuus testiaineistossa oli noin 20%. Näissä tapauksissa paikkatiedosta puuttui kokonaan tai osittain tieto siitä, mihin löydetty sijainti liittyi. Tulosta voitaneen pitää kohtuullisena, kun ottaa huomioon kuvailutietoa etsivän algoritmin yksinkertaisuuden (luku 6.4).

Huolimatta epätarkkuudestaan www-haut toimivat odotetusti paremmin eikaupallisten kohteiden haussa. Muutamissa testin tapauksissa ainoat tulokset saatiin www-haulla. Fonectan palvelussa on puolestaan lisäominaisuus, jonka ansiosta voi saada tietoja kohteesta, joka ei sijaitse samassa kunnassa kuin käyttäjä. Jos esimerkiksi sijaintia vastaavan kunnan alueella ei ole kiinalaista ravintolaa, Fonectan tuloksina palautetaan lähimmän toisella paikkakunnalla olevan kiinalaisen ravintolan tiedot.

Hakutuloksien järjestämisessä käytetään kriteerinä löydetyn kohteen ja käyttäjän välistä etäisyyttä. Etäisyys lasketaan kahden tasokoordinaattipisteen välisenä etäisyytenä. Ohjelman suorittama etäisyyden mittaaminen tarkistettiin satunnaisesti kahden ensimmäisen testitapaustyyppin yhteydessä. Tarkistus tapahtui käyttämällä maanmittauslaitoksen www-palvelua. Palvelussa olevalla etäisyysmittarilla voi mitata kahden sijainnin välistä etäisyyttä [14]. Etäisyyden määrittäminen toimi tarkistetuissa kohteissa oikein.

Kuvassa 13 on autenttinen esimerkki sovelluksen käyttäjälle tuottamasta hakutuloslis-
tasta. Kuvassa näkyvät listan kolme ensimmäistä (lähintä) kohdetta hakusanalla ”pizze-
ria”, kun käyttäjän sijainti on Länsikatu 15 Joensuu. Kohteen etäisyys on ilmoitettu
metereissä.

Kuvaus:	JOENSUU Toimiala Ravintolat Pub-pizzeria Gaude
OSOITE	
Etäisyys:	80.2122185206
Katusosoite:	Länsikatu 18
Postinumero:	80110
Postitoimipaikka:	JOENSUU
Kunta:	joensuu
Linkki:	http://www.yritystele.fi/keyword/header_code/82240/area_code/167/area_name/Joensuu/header_name/Ravintolat
Kuvaus:	puhelin 013 227991 Pub-pizzeria Gaude
OSOITE	
Etäisyys:	80.2122185206
Katusosoite:	Länsikatu 18
Postinumero:	80110
Postitoimipaikka:	JOENSUU
Kunta:	joensuu
Linkki:	http://www.yritystele.fi/keyword/header_code/82240/area_code/167/area_name/Joensuu/header_name/Ravintolat
Kuvaus:	PIZZASPECIAL : puh 013 220246
OSOITE	
Etäisyys:	653.449309434
Katusosoite:	Koskikatu 17
Postinumero:	80100
Postitoimipaikka:	JOENSUU
Kunta:	joensuu
Linkki:	http://www.020202.fi/media/118576

Kuva 13: Sovelluksen käyttäjälle tulostamia hakutuloksia.

7.1.3 Hakuprofilitestin tuloksia

Kolmannessa testitapaustyypissä testattiin hakuprofilien muodostamista. Testissä käy-
tetyt profiilit olivat:

- *Automatka*. Profiili liittyy autolla matkustavan tarpeisiin ja koostuu ”Majoitus”, ”Hinauspalvelu” ja ”Huoltoasema” -hakusanoista.
- *Vapaa-aika*. Hakusanoja jotka liittyvät vapaa-aikaan: ”Kiinalainen ravintola”, ”Elävää musiikkia”.
- *Sekalaiset*. Ei liity mihinkään teemaan, mielivaltaisia hakusanoja: ”Tietokone-
tarvikkeet”, ”Ravirata”, ”Kirjasto”.

Profiileilla tehtiin hakuja neljästä eri sijainnista. Tietoja haetaan sekä [www-sivuilta](#) että Fonectan palvelusta. Taulukossa 9 on esitetty profiilitestin tulokset.

Koska kahdessa aikaisemmassa testitapaustyyppissä tutkittiin ohjelman toimintaa yksittäisillä hakusanoilla, ei tällä kertaa keskitytty tarkemmin löydetyn tiedon laatuun. Testitapausten tarkoituksena oli kokeilla toteutetun profiiliratkaisun käytettävyyttä testikokoonpanolla. Lisäksi useiden hakujen peräkkäinen suorittaminen antaa tietoa ohjelman vakaudesta.

Tulostaulukoissa on laskettu löydettyjen relevanttien ja ei-relevanttien hakutulosten määrä. Hakujen onnistumista kuvaa myös se, onko kaikilla profiilin hakusanoilla saatu tuloksia. Edellisten testitapaustyyppien tapaan on kirjattu myös ohjelman suoritus aika.

Taulukko 9: Hakuprofiileiden testaustuloksia

Käyttäjä Kontiolahdessa				
<i>Profiili</i>	<i>Relevantit</i>	<i>Ei-rel.</i>	<i>Tuloksia kaikilla hauilla</i>	<i>Aika[sek.]</i>
Automatka	14	3	Kyllä	120
Vapaa-aika	6	5	Kyllä	65
Sekalaiset	14	36	Kyllä	35
Käyttäjä Joensuussa				
<i>Profiili</i>	<i>Relevantit</i>	<i>Ei-relevantit</i>	<i>Tuloksia kaikilla hauilla</i>	<i>Aika[sek.]</i>
Automatka	47	29	Kyllä	90
Vapaa-aika	5	1	Kyllä	80
Sekalaiset	16	19	Kyllä	100
Käyttäjä Muhoksella				
<i>Profiili</i>	<i>Relevantit</i>	<i>Ei-relevantit</i>	<i>Tuloksia kaikilla hauilla</i>	<i>Aika[sek.]</i>
Automatka	19	6	Kyllä	125
Vapaa-aika	6	0	Ei	65
Sekalaiset	12	3	Kyllä	100
Käyttäjä Ivalojoella, Inarissa				
<i>Profiili</i>	<i>Relevantit</i>	<i>Ei-relevantit</i>	<i>Tuloksia kaikilla hauilla</i>	<i>Aika[sek.]</i>
Automatka	39	13	Kyllä	80
Vapaa-aika	0	0	Ei	70
Sekalaiset	2	0	Ei	100

Profilin käyttö testikokoonpanossa on miellyttävämpää kuin yksittäisten hakusanojen käyttö. Varjopuolena on ohjelman tuottama suuri tulosten määrä, sillä tulosten läpikäynti pienellä näytöllä vie aikaa. Ohjelman suoritusaika on suoraan verrannollinen profiilin sisältämien hakujen määrään. Testissä suoritusaikojen keskiarvo oli noin 86 sekuntia. Ajat on mitattu tilanteessa, jossa mitään hakuja ei löydetä suoraan tietokannasta. Kun hakujen määrä lisääntyy, joutuu myös toteutuksen vakaus koetukselle. Selainpohjaisena se on jo sinällään altis erilaisille yhteysongelmille. Yhteysongelmien riskiä lisää ajonaikainen tiedonhaku Fonectan ja Maanmittauslaitoksen palvelimilta. Varsinkin viikonloppuisin on maanmittauslaitoksen palvelimilla käyttökatkoksia. Testin aikana ei yhteysongelmia ollut.

7.1.4 Www-hakujen saanti ja tarkkuus

Ideaalinen hakupalvelu tarjoaisi hakijalle hakutuloslistan, joka ei sisällä hyödytöntä tietoa, eikä hukkaa yhtään relevanttia tulosta. Www-sivuille tehdyt haut tuottavat usein turhia tuloksia ja monesti myös epäonnistuvat tärkeiden tietojen löytämisessä. Jotta hakutulosten onnistuneisuutta voitaisiin arvioida, täytyy selvittää myös löytymättä jääneiden relevanttien tulosten määrä.

Relevanttien, epärelevanttien ja löytymättä jääneiden relevanttien tietojen avulla voidaan määritellä haun onnistuneisuutta kuvaavat suureet *saanti* (recall) ja *tarkkuus* (precision). Jos haku tuottaa yhteensä M dokumenttia ja niiden joukosta löydetään P relevanttia tulosta, haun tarkkuus saadaan kaavalla $\frac{P}{M}$. Jos esimerkiksi manuaalisesti tutkimalla saadaan selville, että kaikkien (löydettyjen ja löytymättä jääneiden) relevanttien tulosten kokonaismäärä on R , voidaan haun saanti laskea kaavalla $\frac{P}{R}$. Täydellinen tarkkuus saavutetaan silloin, kun kaikki tulokset ovat relevantteja. Täydellinen saanti puolestaan edellyttää kaikkien tietovarastossa olevien relevanttien tietojen löytämistä [11].

Koesovelluksen tapauksessa laskettiin saanti ja tarkkuus kahden ensimmäisen testitapaustyyppin www-hakujen hakutuloksille. Ennen laskuja tehtiin muutamia oletuksia:

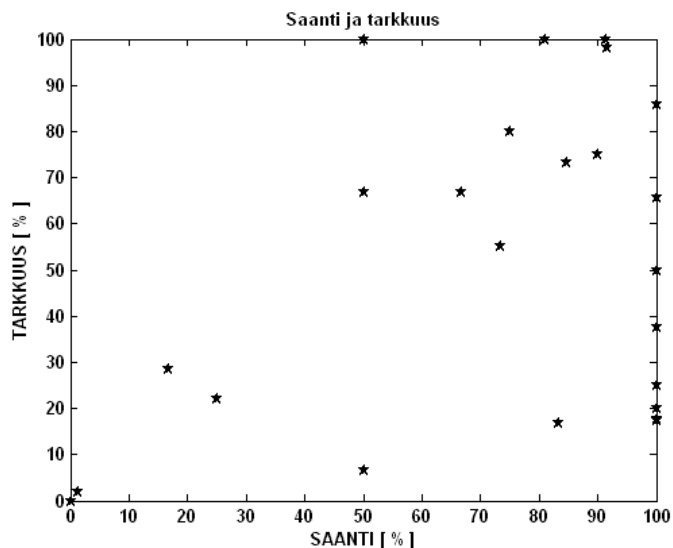
- Kokonaisten dokumenttien sijaan hakuja arvioidaan hakutuloksien perusteella. Www-haun suorittava Google palauttaa oletuksena enintään kymmenen dokumenttia. Arvion kohteena ovat tulokset eristetään näiden dokumenttien sisälöstä ja niiden lukumäärä voi vaihdella suuresti.

- Relevantiksi tulokseksi katsotaan koordinaattien hakemisen mahdollistavan osoitteen ja hakuun liittyvän sisällön omaava tulos.
- Tulos on relevantti vain, jos se sisältää ymmärrettävän kuvausosion. Vaikka hakutulos sisältää kelvollisen osoitteen ja liittyy hakuun, sitä ei katsota relevantiksi jos kuvausosio puuttuu tai se ei ole käyttäjän ymmärrettävissä. Tarkkuuden laskennassa tällainen hakutulos katsotaan epärelevantiksi tulokseksi ja saannin laskennassa löytymättömäksi relevantiksi tulokseksi.
- Kaikki relevantit hakutulokset ovat keskenään samanarvoisia. Kaikki haun kohteena olevaan paikkakuntaan liittyvät relevantit hakutulokset ovat yhtä onnistuneita. Hakutulokset esitetään käyttäjälle järjestyksessä, mutta järjestys riippuu ainoastaan käyttäjän sijainnista.
- Jos relevantteja hakutuloksia ei ole löydettävissä, haun saantia ei voida määritellä.
- Jos haku ei palauta yhtään tulosta, haun tarkkuutta ei voida määritellä.

Hakusovellusten laatua arvioidaan usein saannin ja tarkkuuden välisellä suhteella. Tulokset esitetään sijoittamalla koordinaatistoon saanti/tarkkuus -pisteitä, sekä niiden perusteella interpoloituja käyriä [11]. Kuvassa 14 on koesovelluksen [www-hauille](#) lasketut saanti/tarkkuus -pisteet. Jokainen piste perustuu yhden haun tuottamiin tuloksiin. Pisteiden arvot on saatu seuraavasti:

- M = kaikkien sovelluksen löytämien tulosten yhteismäärä.
- P = sovelluksen löytämien täysin hakua vastaavien ja täydellisiä osoitteita sisältävien tulosten määrä (löydetty relevantit tulokset).
- R = löydettyjen relevanttien ja löytämättä jääneiden relevanttien tulosten yhteismäärä. Löytämättä jääneiden relevanttien tulosten määrä on saatu tutkimalla manuaalisesti samat [www-sivut](#), joita sovellus on tiedonhaussa käyttänyt.
- Tarkkuus: $\frac{P}{M} * 100 \%$, saanti: $\frac{P}{R} * 100 \%$

Jos haku onnistuu pitämään epärelevanttien tulosten osuuden pienenä, tarkkuusprosentti on korkea. Korkea saantiprosentti puolestaan kertoo, että on onnistuttu löytämään suurin osa löydettävissä olevasta relevantista tiedosta. Ideaalisen haun tapauk-



Kuva 14: Kahden ensimmäisen testitapaustyypin testeissä havaitut saanti/tarkkuus - pisteet.

nessä sekä saanti- että tarkkuusprosentit ovat 100. Saanti- ja tarkkuustulos on sitä parempi mitä lähempänä koordinaatiston oikeaa yläkulmaa pisteet sijaitsevat. Kuvassa 14 näkyville koesovelluksen saanti/tarkkuus -pisteille laskettiin myös keskiarvo. Saannin keskimääräinen prosentti oli 67 ja keskimääräinen tarkkuusprosentti 47.

Tuloksia voidaan käyttää vertaillessa sovelluksen suorituskkyä myöhemmin tehtyjen parannusten jälkeen. Vertailu ei kuitenkaan ole ongelmaton, sillä samat haut eivät välttämättä tuota samoja tuloksia. Www on kokoajan muutoksen tilassa ja lisäksi erilaiset yhteys- ja palvelinongelmat vaikuttavat hakutuloksiin.

7.2 Havaittuja ongelmia

Tässä luvussa on esitellään testauksen aikana havaittuja ongelmatilanteita ja puutteita. Muutamissa tapauksissa ongelmat johtuvat suoraan huonosta suunnittelusta, toisinaan taas kyse on käytettyihin tietovarastoihin liittyvistä ominaispiirteistä.

7.2.1 Sijaintitiedon eristämisongelmat

Kun koesovellus yrittää löytää sijaintitietoja www-sivuilta, se aloittaa etsimällä katuosoitteeseen viittaavia sanoja. Löytäessään merkkijonon, jonka loppuosa on ”katu”, ”kuja”, ”polku”, ”raitti” tai ”tie” ohjelma alkaa etsiä kyseisen sanan läheisyydessä mahdollisesti olevia muita osoitteen osia. Menetelmän heikko kohta on ilmeinen: sellaiset osoitteet, joissa ei ole edellä kuvattua loppuosaa, jäävät löytämättä. Jos katuosoite on useamman sanan yhdistelmä, kuten ”Pohjoinen rautatienkatu”, sovellus ei tunnista sanaa ”Pohjoinen” katuosoitteeseen liittyväksi. Toisinaan katuosoitteen sijalla on postilokero (PL) -kenttä. Myös tässä tilanteessa osoitteen eristäminen epäonnistuu.

Kun katuosoitteeseen viittaava sana on löytynyt, ohjelma aloittaa postinumeron, kunnan tai postitoimipaikan etsinnän. Etsintä kohdistuu enintään viiteen katuosoiteehdokkaan *jälkeen* esiintyvään sanaan. Tyypillisten suomalaisten osoitteenkirjoitustapojen kohdalla haku toimiikin. Osoite jää kuitenkin löytymättä jos katuosoite on kirjoitettu postinumeron, postitoimipaikan tai kunnan jälkeen, tai jos eri osoite-elementit eivät sijaitse tekstissä lähekkäin. Molempia virhetilanteita esiintyi testauksen yhteydessä.

Joissakin tapauksissa osoite voi jäädä löytymättä siksi, että tarpeellista tietoa on hävinnyt www-sivun lähdekoodin käsittelyn yhteydessä. Luvussa 6.4 kerrottiin menetelmästä, jolla ylimääräinen muotoilutieto suodatetaan pois tutkittavan tekstin joukosta. Muotoilutagien poistoon liittyy kuitenkin ongelma. Jos osoite-elementti on määritelty sivun tekijän toimesta esimerkiksi linkiksi tai sijoitettu muuten huolimattomasti muotoilutagien sisään, tieto häviää tagien poiston yhteydessä. Koska selaimet kykenevät näyttämään huolimattomalla tai jopa virheellisellä html-kielellä ohjelmoituja sivuja, voi mainitun kaltainen tilanne syntyä myös vahingossa.

Edellä kuvatut ongelmat koskivat www-sivuilta tapahtuvaa osoitteiden eristämistä. Fonetan palveluun kohdistuvien hakujen piti sen sijaan olla ongelmattomia, sillä tulossivut näyttivät ensialkuun vakiomuotoisilta. Osoittautui kuitenkin, että tulossivujen sijaintitietojen esitystapa vaihteli. Suurin osa osoitteista oli normaalimuodossa mutta joukossa oli puutteellisia, ylimääräistä tietoa sisältäviä tai jopa kokonaan puuttuvia osoitekenttiä. Lisäksi kaikkien postilokero (PL) -kenttiä sisältävien osoitteiden eristäminen epäonnistui.

7.2.2 Osoitteet ja koordinaatit

Koesovellus hakee löytämilleen osoitteille koordinaatit ajonaikaisesti maanmittauslaitoksen www-sivulta. Osoitteiden muodostamisen apuna käytetään puolestaan Suomen Postin postinumeroihin perustuvaa kuntaluetteloa. Postin osoitejärjestelmä ei kuitenkaan täsmää kaikissa tapauksissa ohjelmassa käytettyyn maanmittauslaitoksen osoite/koordinaatti -tietokannan kanssa. Esimerkiksi osoitteella Kilvenniementie 80710 Lehmo, on postin järjestelmässä määriteltä kunnaksi sekä Joensuu että Kontiolahti. Koordinaattihaku ei kuitenkaan löydä koordinaatteja jos kuntana on Joensuu. Löytyi myös muita vastaavia tapauksia, jossa yksi postinumero voi vastata useampaa kuin yhtä kuntaa. Eräs ongelmia aiheuttava postinumeroalue testitapauksissa oli Koli. Postin järjestelmässä kunnaksi määrittyy Juuka, Kontiolahti tai Lieksa, mutta koordinaattien hakemiseksi sen tulisi olla Lieksa.

Maanmittauslaitoksen palvelimelle tehtävät koordinaattien haut ovat melko häiriöalttiita. Laitoksen www-sivuilla haetaan koordinaatit jokaiselle löydetylle osoitteelle erikseen. Tällaisia hakuja saattaa tulla ohjelman toimiessa useita satoja. Dokumentoidun testiohjelman aikana yhteyskatkoksia ei tapahtunut, mutta ohjelman suunnitteluvaiheessa tehdyissä testeissä koordinaattien haku epäonnistui muutamia kertoja. On ilmeistä, että tämän tyyppinen toteutus koordinaattien saamiseksi ei ole riittävän vakaa.

7.2.3 Hakutuloksiin liittyviä ongelmia

Käyttäjän kannalta hankalimpia ovat täysin epärelevantit hakutulokset. Paitsi että ne turhaan kasvattavat tuloslistan pituutta, ne myös saattavat johdattaa käyttäjää harhaan. Ongelma on todellinen silloin kun halutulos ei sisällä lainkaan kuvausosiota. Epärelevantteja hakutuloksia syntyy pääasiassa kolmella eri tavalla:

- Www-sivu sisältää yhden tai useamman hyödyllisen tuloksen, mutta sen lisäksi samalta sivulta löytyy täysin erityyppisten palveluiden ja kohteiden tietoja samalta paikkakunnalta. Tyypillisiä ongelmasivuja ovat ”kaikki palvelut kunnasta x” tai ”yritykset alueella x” -tyyppiset sivut.
- Hakuehtoja vastaavaa paikkatietoa löytyy www-sivulta, mutta hakuehdon hakusana esiintyy kyseisellä sivulla eri merkityksessä kuin käyttäjä oli tarkoittanut.

- Hakuehdossa oleva hakusana löytyy sivulta, mutta löydetty hakuehdon mukainen sijaintitieto ei liity mitenkään siihen. Tyypillinen esimerkki on vaikkapa haku sanalla ”Ravirata” Joensuun alueella. Hakutuloksena saadaan Joensuun liikuntatoimiston osoite. Syynä ovat liikuntatoimiston www-sivut, joilla on toimiston osoite mutta joilla esiintyy myös sana ”Ravirata”.

Eräs hakutuloksien onnistuneisuuteen vaikuttava ongelma liittyy käyttäjän sijainnin perusteella tapahtuvaan hakuehdon määrittämiseen. Koesovellus määrittää käyttäjän sijaintikunnan osaksi hakuehtoa. Kunnaksi valitaan se, jonka keskus on lähimpänä käyttäjän sijaintia. Tekniikka toimii melko hyvin harvaan asutulla alueella, missä taajamat ovat pieniä ja kaukana toisistaan. Muualla syntyy kuitenkin ongelmia. Jos käyttäjän sijainti on esimerkiksi Helsingin Malmilla, sovellus tarjoaa hakutuloksia Vantaalta vaikka hakua vastaavia kohteita löytyisi varmasti Malmiltakin. Kuntien ja käyttäjän välisiä etäisyyksiä käsitellään tasopisteiden välisinä etäisyyksinä, joten kuntiin kuuluvien alueiden kokoeroja ei huomioida mitenkään. Näin syntyy tilanne, jossa käyttäjä on Helsingin alueella, mutta lähempänä Vantaan kuin Helsingin keskustaa. Tällöin käyttäjä hyvin todennäköisesti ei löydä lähintä hakemaansa kohdetta. Tätä ongelmaa ja sen mahdollisia ratkaisuja käsittelevät myös Hariharan, Fränti ja Mehta artikkelissaan [10].

7.2.4 Sovelluksen käytettävyyteen ja testilaitteistoon liittyviä ongelmia

Suoritettuaan käyttäjän antamien hakuehtojen mukaiset haut, koesovellus tulostaa hakujen tuloksia sisältävän listan. Jos hakuehdot ovat sopivat, tuloslista on helposti käsiteltävissä myös testilaitteiston pieninäyttöisellä päätelaitteella. Esimerkkinä toisenlaisesta tapauksesta oli ensimmäisen testitapaustyyppin Helsingin alueella tapahtunut haku hakusanalla ”autokorjaamo” (luku 7.1.1), jolloin hakutuloksia kertyi peräti 114. Tällaisen listan vierittäminen pienellä ruudulla ei ole enää järkevää. Vaikka käyttäjä keskittyneekin useimmiten listan alussa oleviin lähimpiin kohteisiin, olisi käytettävyyden kannalta hyvä rajoittaa käyttäjälle oletusarvoisesti näytettävien hakutulosten määrää.

Osaltaan hakutuloslistaa kasvattavat siinä olevat duplikaatit. Www-sivuilta tietoja haettaessa sama kohde saattaa löytyä useammalta sivulta. Lopullisessa tuloslistassa kohde esiintyy useampaan kertaan toisistaan eroavilla kuvausosiolla varustettuna. Toisinaan tästä on hyötyäkin. Esimerkiksi hakutuloksen puutteellinen kuvausosio voi käyttäjän kannalta täydentyä toiselta www-sivulta löytyneestä, saman kohteen paremmin onnis-

tuneesta kuvauksesta.

Testilaitteiston päätelaitteena toiminut Nokia 6630 -puhelimien selainohjelma ei kyennyt käyttämään mitään Fonectan palvelusivuja. Ongelma ilmeni kun laitteella yritettiin seurata sovelluksen löytämien Fonectan hakutulosten medialinkkejä. Asiaa selvitellessä kävi ilmi, että ainoastaan Nokian Communicator-sarjan puhelimilla linkkien avaaminen onnistuu.

Testikokoonpanossa paikannusvälineenä toiminut GPS-laite asettaa sovelluksen käytölle rajoitteita. GPS ei yleensä toimi sisätiloissa. Joissakin rajatapauksissa laite saattoi toimia esimerkiksi ikkunan lähellä. Autokäytössä laitteen sai toimimaan sijoittamalla se kojelaudan päälle. Sisätilarajoituksia pystyy jonkin verran kiertämään laitteen sisäisen muistin avulla. Sen ansiosta tallessa ovat viimeiset päivitettyt koordinaatit ennen signaalin katkeamista. Tällöin sovellusta voidaan käyttää ja käyttäjän sijainniksi saadaan koordinaatit rakennuksen ulkopuolella. GPS-paikantimien eräs yleinen ongelma on suurehko virran kulutus. Testilaitteen pieni koko asettaa rajoituksia akun koolle, joten laitetta pystyy käyttämään yhtäjaksoisesti vain muutamia tunteja.

7.2.5 Muita ongelmia

Tämänhetkisessä muodossaan koesovelluksen tiedonhakuun liittyy lukuisia rajoitteita. Osa on tarkoituksellisesti asetettu estämään toiminnan liiallista hidastumista, osa taas johtuu toteutustavoista. Rajoitusten seurauksena hakutuloksia jää löytämättä.

Ohjelman käyttämän Google-haun tulossivujen määrä on rajoitettu kymmeneen sivuun. Googlen ranking pyrkii huolehtimaan siitä, että olennaisimmat tulossivut ovat ensimmäisinä. Jos haetaan vaikkapa pizzeriaa suuren kaupungin alueelta, relevanttia tietoa sisältäviä sivuja voi olla useita kymmeniä. Tällöin hyödyllisiä hakutuloksia jää löytymättä tulossivumäärän rajoittamisen seurauksena.

Etsiessään tietoa www-sivuilta ohjelma tyytyy tutkimaan vain Googlen löytämät hakuehtoa vastaavat linkit. Usein kuitenkin tämä linkki on vain jonkin sivuston etusivu. Esimerkiksi hakusanalla ”Ravintola” sijainnilla ”Joensuu”, ohjelma saa tutkittavakseen www.joensuuopas.com -sivuston ravintolat-sivun. Se sisältää Joensuulaisten ravintoloiden nimiä ja kuvauksia, mutta ei lainkaan sijaintitietoa. Klikkaamalla ravintolan linkkiä, aukeaa sivu josta tarkka osoite selviää. Koesovellus ei kuitenkaan seu-

raa linkkejä, joten hakutuloksia ei tältä sivustolta löydetä. Samankaltainen ongelma esiintyy myös kehyksiä käyttävillä sivuilla. Sivuston pääsivu ei yleensä sisällä haluttua tietoa, vaan tiedon löytämiseksi ohjelman tulisi seurata kehyksiin osoittavia linkkejä.

Toisinaan yrityksen osoite- ja yhteystiedot on esitetty sivustolla olevassa kuvassa, joten lähdekooditekstiä tutkiva ohjelma ei tämänkaltaista informaatiota pysty hyödyntämään.

Fonectan palveluhakua käyttäessään ohjelma tutkii enintään kymmenen ensimmäistä tulosta. Tähän on päädytty siksi, että tulossivujen lataaminen ja tutkiminen on hidasta. Jos haku kohdistuu johonkin tyypilliseen palveluun suuressa kaupungissa, tuloksia löytyy todella paljon. Esimerkiksi hakusanoilla ”ravintola” ja ”helsinki” löytyy 1335 kohdetta (136 tulossivua). Koska koesovellus joutuu tulossivujen lataamisen ja tutkimisen lisäksi hakemaan jokaisen osoitteen koordinaatit erikseen Karttapaikan www-sivuilta, suuret kohdemäärät hidastavat ohjelman toimintaa huomattavasti. Onnistuneen Fonecta-haun kaikki tulokset ovat kuitenkin relevantteja. Niinpä edellä kuvattu hakutulosten määrän rajoittaminen voi heikentää ohjelman käyttökelpoisuutta. Hakutulosten esitysjärjestyksessä ei ole mitään havaittavaa säännönmukaisuutta, joten käyttäjän kannalta tärkein kohde voi jäädä rajauksen ulkopuolelle.

Hakutuloksien määrään ja laatuun vaikuttaa myös palvelimien ja verkon toiminta. Eri vuorokaudenaikoina tai viikonpäivinä saadaan erilaisia tuloksia. Esimerkiksi koordinaattien hakua haittaavat maanmittauslaitoksen palvelimien käyttökatkokset. Näitä esiintyy erityisesti viikonloppuisin. Tietoliikenteen ruuhkaisuus verkossa voi vaikuttaa siten, että joillekin www-sivuille ei päästä lainkaan. Jos palvelin ei vastaa tietyn aikarajan kuluessa, yhteysyritys lopetetaan ja siirrytään yrittämään seuraavaa. Edellä kuvatusta seuraa myös eräs hakutulosten tallentamiseen liittyvä ongelma. Kun käyttäjä on tyytyväinen hakutuloksiin, hän voi halutessaan tallentaa ne tietokantaan. Jos käyttäjä hakee myöhemmin samoilla hakuehdoilla, ohjelma palauttaa tietokannassa olevat tulokset uuden haun sijasta. Mikäli käyttäjän ensimmäinen haku tapahtui huonoissa olosuhteissa, tietokantaan ei ehkä tallentunut paras mahdollinen aineisto.

Hakuprofiilia muodostettaessa käyttäjä tallentaa mielestään onnistuneet hakutulokset tietokantaan. Samalla tallentuvat kyseisen haun hakusanat automaattisesti profiliin. Kun käyttäjä käyttää ohjelmaa valmiilla profiililla, hakutulokset tallennetaan automaattisesti tietokantaan. Toteutuksen pohjana oli ajatus, että kertaalleen hyväksi havaittu hakusana tuottaa jatkossakin hyviä tuloksia, eikä käyttäjän tarvitse enää päättää tallenne-

taanko hakutulokset vai ei. Käytännössä käy kuitenkin niin, että samoilla hakusanoilla mutta eri paikkakunnilla suoritettut haut poikkeavat sekä määrällisesti että laadullisesti huomattavan paljon toisistaan. Tämän johdosta tietokantaan tallentuu myös huonoja hakutuloksia.

Ohjelman tietovarastona käyttämiin www-palveluihin liittyy muutamia epävarmuustekijöitä. Toteutusvaiheessa sekä Googlen että Fonectan palvelusivut muuttuivat kaksi kertaa. Näiden seurauksena ohjelma lakkasi useaan otteeseen toimimasta. Esimerkiksi Fonectan tapauksessa oli pakko ohjelmoida uudelleen kaikki siihen liittyvät aliohjelmat. Tiedossa ei ole johtuivatko muutokset tarpeellisista päivityksistä vai halusta vaikeuttaa ulkopuolisten mahdollisuuksia käyttää palveluita ohjelmallisesti. Joka tapauksessa on todennäköistä, että vastaavia ongelmia esiintyy jatkossakin.

8 Kehitysideoita

Tässä luvussa keskitytään siihen, miten koesovellusta tai sen kaltaisia muita sijaintitietoa hakuehtojen osana hyödyntäviä sovelluksia voitaisiin kehittää. Käsittelyn pohjana ovat koesovelluksen testauksessa havaitut ongelmat ja sen käytöstä saadut kokemukset.

8.1 Tiedonhaun tehostaminen

Kuten luvussa 7.2.5 todettiin, koesovellus etsii tietoa Google-haun palauttamilta sivuilta, mutta ei kykene seuraamaan näillä sivuilla olevia, mahdollisesti hyödyllisiä jatkolinkkejä. Usein Googlen palauttama linkki johtaa vain sivuston etusivulle, jossa esitellään sivuston sisältöä. Ainakin suomalaisilla palvelusivustoilla hyödyllistä tietoa löytyy usein heti seuraavan tason linkistä. Seuraavanlainen menettely voisi lisätä hyödyllisten hakutulosten lukumäärää:

- Tutkitaan alkuperäisen linkkilistan osoittama sivu normaalisti.
- Kerätään sivulta saman domainin alaiset linkit talteen.
- Jos alkuperäiseltä sivulta ei löydy sijaintitietoa, tutkitaan kerättyjen linkkien osoittamat sivut.

Jotta kaikki tärkeä tieto löydetäisiin, täytyy Fonectan palvelusivuilta kerätä kaikki haakuun liittyvät tulokset. Tämä tarkoittaisi nykyisellä sovelluksella kymmenien, jopa satojen tulossivujen tutkimista ja siten ohjelman suorittamiseen käytettävän ajan moninkertaistumista. Parannetussa sovelluksessa tulisikin olla pääsy suoraan palveluntarjoajan tietokantaan. Silloin ei tarvittaisi häiriöaltista ja aikaavievää erillisten tulossivujen lataamista. Sovelluksen käyttöön tarvittavien tietojen kerääminen olisi tietokantahaakuina luotettavaa ja yksinkertaista verrattuna www-sivun lähdekoodista tapahtuvaan tiedon eristämiseen.

8.2 Sijaintitiedon eristäminen www-sivuilta

Sijaintitiedon etsimisen menetelmät www-sivuilta kaipaavat myös parantelua. Nykysovelluksessa kyetään löytämään vain ne osoitteet joissa esiintyy merkkijono ”katu”,

”polku”, ”kuja”, ”raitti” tai ”tie”. Näin ollen esimerkiksi *Itäranta 1, 80100 Joensuu* tyyppinen osoite jää löytymättä. Kenties olisi mahdollista luoda hieman laajempi otos tyyppillisistä lähiosoiteiden päätteistä. Tämän pienen tietokannan avulla voitaisiin tunnistaa tarkemmin osoitteita suomalaisesta tekstistä. Samalla olisi varauduttava tutki-
maan tekstiä myös sen varalta, että osoite ei ole kirjoitettu tyyppillisessä järjestyksessä *katuosoite, postinumero, postitoimipaikka, kunta*. Jos käytössä olisi osoitetietokanta, voitaisiin sitä mahdollisesti käyttää apuna. Tietokantahaku voidaan rajata käyttäjän sijaintikuntaan kuuluviin osoitteisiin. Näihin osoitteisiin vertailemalla voitaisiin muodostaa osoitteita tekstistä löytyneistä osoite-elementeistä.

Toinen tapa osoitetietokannan hyödyntämiselle osoitteiden löytämiseksi voisi olla seuraavanlainen:

- Etsitään tekstistä viisinumeroista merkkijonoa. Jos sellainen löytyy, tarkistetaan tietokannasta onko se kelvollinen postinumero (näin menetellään myös nykysovelluksessa).
- Osoitetietokannasta haetaan postinumeroalueen katuosoitteet.
- Verrataan postinumeron lähiympäristöstä löytyviä sanoja tietokannasta saatuihin postinumeroalueen katuosoitteisiin. Jos löytyy vastaavuutta, muodostetaan tie-
doista osoite.

8.3 Hakutulosten laatu ja määrä

Nykyisessä muodossaan sovellus tulostaa kaikki löytämänsä hakutulokset käyttäjän nähtäväksi. Joissakin tilanteissa hakutuloksia tulee ohjelman käytettävyyden kannalta liian paljon. Joukossa on myös sellaisia tuloksia, joiden hyöty käyttäjälle on vähäinen tai olematon.

Löydetyn paikkatiedon hyödyllisyyttä käyttäjälle on vaikea arvioida ohjelmallisesti. On kuitenkin mahdollista luoda sääntöjä, joiden avulla voidaan jollakin todennäköisyydellä löytää joukosta parhaat ehdokkaat. Nykysovellus järjestää hakutulokset pelkästään etäisyyden perusteella. Seuraavassa vaiheessa voitaisiin muuttaa rankingia siten, että vain todennäköisesti relevantit *ja* lähimmät kohteet ovat listan kärjessä. Jos kohde on lähin mutta todennäköisesti epärelevantti, se sijoitetaan listaan vasta relevanttien tulosten jälkeen. Hakutulos määriteltäisiin relevantiksi silloin kun:

- Hakutulos on löytynyt Fonectan tietokannasta. Jos Fonecta-haku tuottaa osumia, on hakutulos suurella todennäköisyydellä relevantti.
- Hakutuloksen kuvausosio sisältää hakusanoja. Www-hausta saatu tulos on tässä tapauksessa todennäköisesti relevantti, mutta pienemmällä todennäköisyydellä kuin Fonectan hakutulos.
- Hakutulos sisältää kuvausosion, mutta kuvausosio ei sisällä hakusanoja. Tällainen hakutulos määritellään vähemmän relevantiksi kuin kaksi edellistä, mutta kuuluu vielä hyväksyttävien hakutulosten joukkoon.

Listan viimeinen kohta on ongelmallinen. Vaikka hakutuloksen kuvausosio ei sisältäisikään hakusanoja, hakutulos voi olla yhtä onnistunut kuin Fonectan hakutulokset. Toisaalta hakutulos voi olla myös täysin epärelevantti. Myös listan toinen kohta sisältää virhemahdollisuuden. Vaikka kuvausosio sisältäisikin hakusanoja, ne voivat esiintyä tässä yhteydessä toisessa merkityksessä. Todennäköisesti menetelmä kuitenkin parantaa käyttäjän mahdollisuuksia saada näkyviin lähimmät *sekä* relevantteimmat kohteet.

Edellisen listan ulkopuolelle jääneet hakutulokset sijoitetaan tuloslistaan etäisyysjärjestyksessä muiden jälkeen. Tämän jälkeen poistetaan listasta mahdolliset duplikaatit. Eräs tapa määritellä hakutulos duplikaatiksi voisi olla vertailla kahden hakutuloksen etäisyyttä käyttäjästä. Jos etäisyydet ovat täsmälleen samat, on kysymyksessä todennäköisesti sama kohde. Tätä menetelmää käytetään jo nykysovelluksessa poistamaan Fonecta-haun ja www-haun tuloksien päällekkäisyyttä.

Kun hakutuloslista on edellä kuvatuilla tavoilla editoitu, se näytetään käyttäjälle. Oletusarvoisesti tuloksia voitaisiin näyttää vaikkapa 10. Lisää tuloksia käyttäjä voisi saada näkyville niin halutessaan. Kun käyttäjältä kysytään hänen halukkuuttaan tallentaa tulokset tietokantaan, tulisi hänellä olla mahdollisuus heikkolaatuisten hakutulosten poistamiseen.

8.4 Palvelimella sijaitsevat tietovarastot

Koesovelluksessa eräs epävarmuustekijä on osoitteita vastaavien koordinaattien haku verkon yli. Yhteysskatkot ja verkon ruuhkaisuus vaikuttavat suoraan koko ohjelman toimintaan. Eräs ratkaisu voisi olla valmiin osoite- ja koordinaattiaineiston lunastaminen

ohjelman käyttöön. Aineiston sijoittaminen sovelluksen kanssa samalle palvelimelle nopeuttaisi ja vakauttaisi ohjelman toimintaa.

Palvelimelle voitaisiin sijoittaa myös mobiililaitteilla katseltavaksi soveltuvaa kartta-aineistoa. Kun käyttäjä tutkii hakutuloksia, hän voisi valita muutamia kohteita kartalla näytettäväksi. Nämä kohteet käyttäjä näkisi karttapisteinä, kuvailutiedon kera. Vielä hyödyllisempää olisi käyttää jotakin valmista navigaatio-ohjelmistoa. Ne sisältävät valmiiksi osoite/koordinaatti -tietokannat sekä kartta-aineiston. Sovelluksen käyttäjä valitsisi jälleen mieleisensä hakutulokset, joiden koordinaatit välitetään navigaatio-ohjelmalle. Käyttäjä saisi kohteet näkyviin kartalle ja sen lisäksi reittiohjauksen perille niin halutessaan.

Sovelluksen käyttäjien keräämä hakutulostietokanta voisi toimia kaikkien käyttäjien yhteisenä tietovarastona. On todennäköistä, että jossain vaiheessa useammat käyttäjät hakevat tietoa samoilla hakuehdoilla. Jos tietokanta on kaikkien käytössä, hakutulokset löytyvät nopeasti tietokannasta. Jos sovelluksella olisi erittäin paljon käyttäjiä, voisi olla mahdollista kerätä myös omaa osoite/koordinaatti -tietokantaa hakutuloksista. Koesovellus tallentaa nykymuodossaan hakutulostietokantaan myös kohteen koordinaatit.

8.5 Käyttäjän sijaintikunnan määrittäsongelma

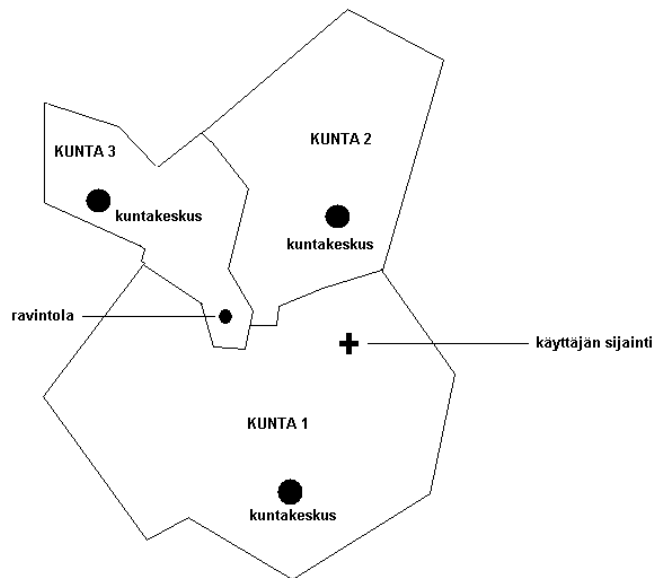
Osoite/koordinaatti -tietokannan avulla olisi mahdollista yrittää ratkaista luvussa 7.2.3 kuvattua ongelmaa. Koesovellus määrittää käyttäjän sijaintikunnan vertailemalla käyttäjän etäisyyttä kuntakeskuksista. Tällöin suuren kunnan laitamilla käyttäjän sijaintikunnaksi saattaakin määrittyä naapurikunta. Jos käytössä olisi osoite/koordinaatti -tietokanta, voitaisiin käyttäjän sijaintikoordinaatteja vertailla tietokannan koordinaatteihin. Lähintä löydettyä koordinaattipistettä vastaavasta osoitteesta saataisiin selville käyttäjän todellinen sijaintikunta. Toisaalta ohjelman nykyinen toimintapa voi olla toimivampi haja-asutusalueella. Suurten, harvaan asuttujen kuntien rajamailla ei useinkaan ole palveluita. Kun käyttäjän sijaintikunnaksi määritellään naapurikunta, jonka keskus on lähempänä, haku voi tuottaa parempia tuloksia.

Sekä taajama- että haja-asutusalueilla toimiva ratkaisu voisi olla seuraavanlainen:

- Osoite/koordinaatti -tietokannan avulla etsitään käyttäjän sijaintia lähimpänä oleva osoite. Osoitteesta selviää sijaintikunta.

- Etsitään lähin kuntakeskus kuntakeskuksien sijainteja sisältävästä tietokannasta.
- Jos osoitteesta löydetty kunta on sama kuin lähin kunta, käytetään hakuehtona vain sitä.
- Jos edellinen ehto ei täyty, suoritetaan haut molempia kuntia käyttäen. Tulokset esitetään etäisyysjärjestyksessä.

Edellä kuvattu ratkaisu parantaa onnistumismahdollisuuksia, mutta täydellinen se ei ole. Jos käyttäjän sijainti on esimerkiksi kolmen kunnan aluerajojen läheisyydessä, voi kunnan määrittäminen jälleen epäonnistua. Ongelmaa on havainnollistettu kuvassa 15. Käyttäjän sijaintikunta on Kunta 1 ja lähimpänä sijaitseva kuntakeskus on Kunta 2. Käyttäjä haluaisi löytää lähimmän ravintolan, mutta se sijaitseekin Kunta 3:n alueella.



Kuva 15: Käyttäjän sijaintikunnan määrittäsongelma

Yksi ratkaisu voisi olla säilyttää nykyinen sijaintikunnan määrittäminen oletuksena ja lisätä mahdollisuus syöttää haun kannalta tarkoituksenmukaisin kunta ohjelmaan käsin. Tämä lisäominaisuus olisi muutenkin hyödyllinen. Käyttäjä voisi halutessaan hankkia hakutuloksia mistä tahansa kunnasta, omasta sijainnistaan riippumatta.

8.6 Maariippumattomuus

Koesovellus on suunniteltu toimimaan yksinomaan Suomessa. Haut on rajattu suomenkielisille web-sivuille ja apuna käytetään suomalaisia palveluntarjoajia. Jotta ohjelmasta olisi todellista hyötyä, tulisi jatkokehityksessä pyrkiä mahdollisimman suureen maariippumattomuuteen. Maariippumattomuus on kuitenkin haastava tavoite. Seuraavassa muutamia ongelmakohtia:

- Kieliongelma. Sovellus hakee tietoa www-sivuilta, joten tiedon löytämiseksi haut on suunniteltava huomioimaan kielialueet.
- Koordinaattijärjestelmien ja koordinaatistojen erot on huomioitava toteutuksessa.
- Osoitejärjestelmien erot vaikeuttavat esimerkiksi osoitetiedon eristämistä www-sivuilta.
- Eri maissa on erilaisia palveluntarjoajia.

Jos oletetaan, että ohjelmaa käyttävät eri maissa vain kyseisen maan kansalaiset, voidaan maariippumaton sovellus toteuttaa varsin suoraviivaisesti. Palvelimille voitaisiin esimerkiksi sijoittaa maakohtaiset osoite-, koordinaatti- ja paikkatietokannat. Samoin huomioitaisiin myös maan omat palvelusivustot. Www-haut rajattaisiin vain saman kielialueen sivustoille. Hakutulokset palautettaisiin maan omalla kielellä. Käyttäjän sijaintikoordinaattien perusteella ohjelma päättelisi mitä maakohtaista ohjelmapakettia käytetään.

Toteutus jossa käyttäjä käyttää ohjelmaa liikkueessaan maasta toiseen, on huomattavasti haastavampi. Www on tavattoman suuri ja kaikin tavoin järjestäytymätön tietovarasto. Jos käyttäjä haluaa hakea ohjelmalla palveluita ja kohteita eri maissa, hän törmää kieliongelmaan.

Otetaan esimerkiksi suomalainen, kohtuullisesti englantia taitava turisti, joka matkustaa Ranskaan. Päästyään perille hän haluaa selvittää lähimpien ravintoloiden sijainnin koesovelluksella. Miten haku tulisi suorittaa? Jos ohjelma hakee tietoa ranskan kielellä, pitäisi käyttäjän ensinnäkin osata antaa ranskankieliset hakusanat. Jos käyttäjä kykenee antamaan oikeat hakusanat, haku tuottaa tuloksia. Paikkatiedon eristäminen

onnistuu maakohtaisen tietokannan avulla. Käyttäjän on kuitenkin vaikea arvioida hakutuloksien onnistumista jos kuvaus on ranskaksi. Hän voi vain arvailla liittyykö tulos hakusanaan. Jos haku suoritetaan englannin kielellä, tuloksia saadaan ehkä vain tyyppillisistä kansainvälisistä tai turisteille tarkoitetuista palveluista ja kohteista. Tällöin käyttäjä kuitenkin pystyy paremmin arvioimaan hakujen onnistumista.

Kun ohjelma hakee tietoa maakohtaisilta palveluntarjoajilta, tulosten voidaan olettaa olevan hakusanoja vastaavia. Tässäkin tapauksessa käyttäjän tulee osata antaa hakusana ranskaksi. Lisäksi auttava ranskankielen taito on tarpeen hakutulosten ymmärtämisessä.

Lisää ongelmia on luvassa kun siirrytään maihin, joissa käytetään länsimaista eroavia kirjoitusmerkkejä. Tässä tapauksessa helpoin vaihtoehto on suorittaa haut englanninkielisiltä www-sivuilta.

Nykymuotoisessa webissä edellä kuvatut ongelmat ovat hankalia ratkaista tyydyttävästi. On olemassa pyrkimyksiä, joilla www:ssä olevaa tietoa pyritään järjestämään siten, että tietoa voitaisiin löytää pikemminkin merkityksien kuin yksittäisten hakusanojen perusteella. Tällainen kehitys vähentäisi myös kieleen liittyviä ongelmia, sillä tieto luokiteltaisiin yleisesti käytössä olevien standardien mukaisesti. Tätä asiaa on käsitelty tarkemmin luvussa 9.2.

9 Hakutuloksien parantamismahdollisuuksia

Sijaintitietoa hyväkseen käyttävät kaupalliset hakupalvelut ovat tätä kirjoitettaessa jo arkipäivää. Koesovelluksen kantavana ajatuksena oli mahdollistaa sijaintiin liittyvät haut myös vapailta www-sivuilta. Teoriassa siten olisi mahdollista suorittaa hakujamista tahansa kiinnostuksen kohteesta. Nykyisillä hakumenetelmillä tulokset jäävät käytännössä melko vaatimattomiksi.

9.1 Nykymuotoinen web

Www on koko ajan kasvava ja monimuotoinen tietovarasto. Relevantin tiedon löytämistä helpottamaan on kehitetty erilaisia hakukoneita, mutta niidenkään avulla ei

tiedonhaku aina sujuu toivotulla tavalla. Seuraavassa on lueteltu muutamia www-tiedonhaun haasteita [8]:

- Web on yksinkertaisesti liian suuri tehokkaan tiedonhaun toteuttamista ajatellen. Tietomäärä on satoja teratavuja ja kasvaa nopeasti koko ajan.
- Web-sivut ovat tavallisiin tekstidokumentti-kokoelmiin verrattuna monimuotoisempia. Sivuilla ei ole yhdenmukaista rakennetta. Ne eivät ole missään tiettyssä järjestyksessä eikä niitä voida hakea tehokkaasti otsikoilla tai tekijätiedoilla kirjaston kirjahaun tapaan.
- Www on jatkuvasti muuttuva tietovarasto. Sen lisäksi että tiedon määrä kasvaa koko ajan, myös jo olemassa oleva tieto muuttuu päivitysten seurauksena.
- Www-hakujen käyttäjiä on paljon. Käyttäjien kiinnostuksen kohteet ja tiedolliset valmiudet eroavat toisistaan suuresti.
- Ainoastaan pieni osa www:n informaatiosta on todella relevanttia ja hyödyllistä. On sanottu, että 99% www:n sisällöstä on hyödytöntä 99%:lle www:n käyttäjistä. Yksittäisen käyttäjän kiinnostuksen kohteet edustavat häviävän pientä osaa koko sisällöstä ja ovat näin vaarassa hautautua ei-kiinnostavan tietomassan alle.

Monet hakukoneet käyttävät tiedonhaun perustana avainsanoihin perustuvia hakemistoja [8]. Niiden avulla voidaan löytää se sivujoukko, joka sisältää tietyt avainsanat. Jos käyttäjä onnistuu hakusanojen valinnassa, haku tuottaa tuloksia. Monesti hakutuloksia on hyvin paljon, joukossa myös runsaasti hakijan kannalta käyttökelvottomia dokumentteja. Tämä ongelma johtuu usein siitä, että samalla sanalla on useita eri merkityksiä eri asiayhteyksissä. Toinen ongelma aiheutuu hakusanan synonyymeistä. Tällöin käyttäjän kannalta relevantteja dokumentteja jää löytymättä, koska niissä on käytetty samaa asiaa tarkoittavia, mutta hakusanoista eroavia sanoja ja ilmauksia.

Hakusanoihin liittyvää väärin tulosten ongelmaa voitaisiin helpottaa tutkimalla, esiintyvätkö jotkin erilliset sanat säännöllisesti samoissa dokumenteissa. Tällaisissa tapauksissa voitaisiin johtaa yhteyssääntöjä: esimerkiksi sana *pörssikurssi* merkinnee, että suurella todennäköisyydellä dokumentissa esiintyvät myös sanat *euro*, *osake*, *salkku*. Jos käyttäjä on kiinnostunut sijoittamiseen liittyvistä sivustoista, hakutuloksista poimitaan vain ne sivut, jotka toteuttavat yhteyssäännön.

Www-sivuja voitaisiin myös luokitella automaattisesti. Tässä menetelmässä käytetään etukäteen luokiteltuja dokumentteja ohjelman opetusaineistona. Esiluokitellut dokumentit analysoidaan, jotta niistä voidaan johtaa luokittelumalli. Luokittelumallia hienosäädetään testaamalla sitä muulla aineistolla. Valmista mallia voi tämän jälkeen käyttää uusien dokumenttien automaattiseen luokitteluun.

Hakusanojen suhteen onnistuneidenkin tulossivujen joukossa voi olla kehnolaatuista tai muuten arvotonta materiaalia. Sivujen laatua voidaan yrittää mitata tutkimalla sitä, kuinka monelta ja kuinka tärkeältä sivulta kyseiseen dokumenttiin on viitattu. Näiden ”suositusten” perusteella voidaan hakua vastaavat sivut asettaa järjestykseen. Tämän-tyyppinen menetelmä on käytössä Google hakukoneessa.

Paikkatiedon löytäminen www-sivuilta vaatii nykyisellään sivujen koko sisällön läpikäymistä. Ongelma ei ole niinkään www:ssä itsessään, vaan sinne sisältöä tuottavissa tahoissa. Paikkatietoa hyödyntävien sovellusten kannalta olisi toivottavaa, että paikkatiedon esitystapa ja -paikka dokumenteissa noudattelisi jotakin yleistä standardia. Aiemmin luvussa 4 esiteltyjen geo- ja address-tagien asianmukainen käyttö nopeuttaisi ja tarkentaisi sovellusten toimintaa.

9.2 Semanttinen web

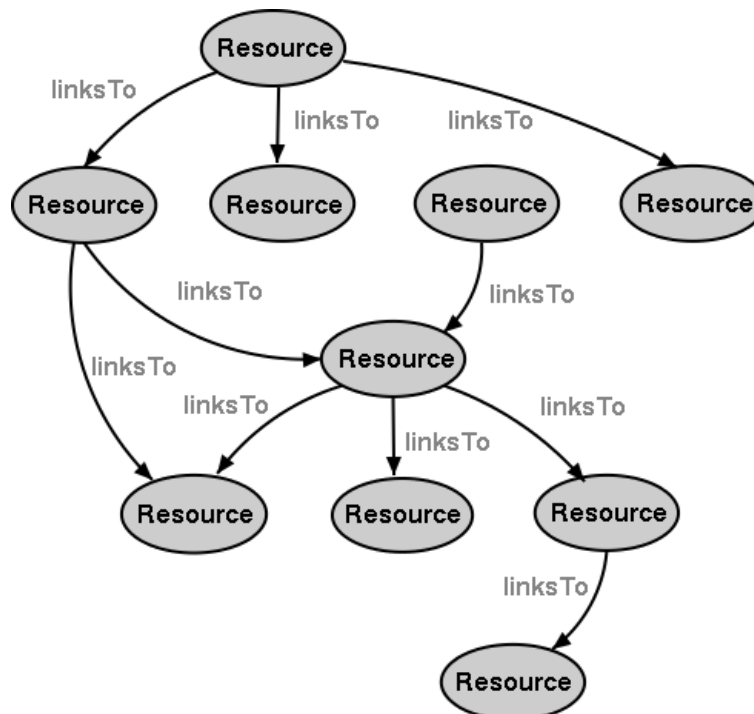
Kuten edellä on havaittu, www:n sisältämän valtavan tietomäärän tehokas hyödyntäminen on nykyisillä menetelmillä vaikeaa. Html-kieltä käytetään lähinnä vain sivujen ulkoasun muotoiluun, eikä sen perusteella voi tehdä juurikaan päätelmiä dokumenttien sisällöstä. Hakusanoilla hakeminen ei onnistu aina toivotulla tavalla, koska samoilla sanoilla on eri merkityksiä eri yhteyksissä. Toisaalta useat eri sanat voivat merkitä samaa asiaa. W3C on aktiivisesti kehittämässä järjestelmää, jossa tietoa tallennettaessa määritellään sanojen merkityksiä ja tietojen välisiä suhteita [28]. Tätä järjestelmää kutsutaan *semanttiseksi webiksi* (semantic web).

Semanttisessa webissä tietoresurssit sisältävät tietoa itsestään [12]. Tämänkaltaista tietoa kutsutaan *metadataksi*. Metadatan tulee olla sellaisessa muodossa, että se on helposti ohjelmallisesti käsiteltävissä. Metadatan malli ja semanttisen webin ”sydän” on *RDF* (Resource Description Framework). RDF:n peruselementti on lausuma. Lausumassa resurssi (*subjekti*) on linkitetty toiseen resurssiin tai arvoon (*objekti*) jonkin kolmannen resurssin (*predikaatti*) kautta.

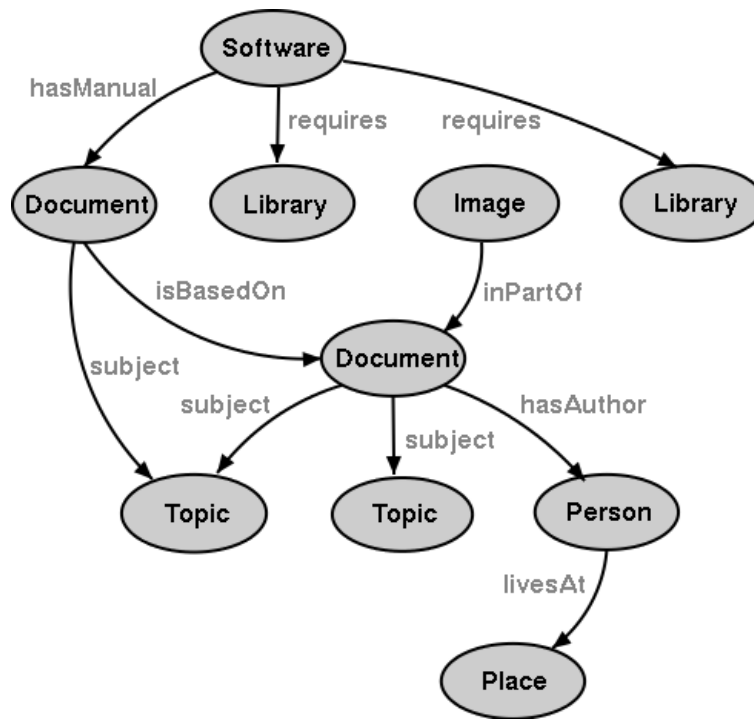
Resurssi (Resource) on mikä tahansa kohde mikä voidaan identifioida *URI:n* (Uniform Resource Identifiers) avulla. Jos tämä dokumentti sijaitisi kuvitteellisen `www.omadomain.fi` -palvelimen juurihakemistossa `gradu.pdf` -nimisenä tiedostona, olisi ”`http://www.omadomain.fi/gradu.pdf`” silloin RDF-resurssi. Tässä tapauksessa URI on `www`-osoite, mutta URI voi identifioida myös kohteita, joihin ei ole tietoverkkoyhteyttä. URI voidaan luoda viittaamaan esimerkiksi henkilöön, yritykseen, kirjaan tai vaikkapa johonkin ei fyysiseen kohteeseen.

RDF-lausuma voidaan ajatella verkkona tai verkon osana. Subjektit ja objektit ajatellaan solmuiksi ja predikaatit suunnatuiksi kaariksi subjektista objektiin. Lausumat yhdistävät `www`-dokumenteissa olevat tiedot toisiinsa. Lausumien avulla selviävät myös asiayhteydet ja tietojen väliset suhteet. Näin tavallisesta, erillisistä dokumenteista muodostuva web voisi muuntua tietoja, yhteyksiä ja merkityksiä kuvaavaksi semanttiseksi webiksi. Kuvissa 16 ja 17 on havainnollistettu nykyisen ja semanttisen webin eroavaisuuksia.

Jos semanttinen web toteutuu visioiden mukaisesti, `www:n` hyödyntäminen tietovarastona tehostuu. Nykyiset hakukoneet perustavat toimintansa `www`-dokumentteja ”sokeasti” läpikäyvään raakaan laskentatehoon. Semanttisessa webissä voitaisiin käyttää



Kuva 16: Wwv nykymuodossaan [17].



Kuva 17: Semanttinen web [17].

älykkäämpiä, asiayhteyksiä ja merkityksiä tutkivia hakukoneita. Esimerkiksi tässä tutkielmassa käsitellyn koesovelluksen toimintaa voitaisiin parantaa seuraavilla tavoilla:

- Hakuehdossa olevia hakusanoja käsiteltäisiin vain käyttäjän tarkoittamassa merkityksessä.
- Sovelluksen ei enää tarvitsisi käydä dokumentteja läpi rivi riviltä sijaintitietoa etsiessään, vaan tieto olisi löydettävissä hakukohteen sijaintia kuvaavasta erillisestä resurssista.
- Haku voidaan rajata palauttamaan vain ne dokumentit, jotka sisältävät sijaintitietoa.

Toteutuakseen semanttinen web tarvitsee asiasta innostuneita sisällöntuottajia ja organisaatioita. Lisäksi tarvitaan yhteisesti sovittuja, määriteltyjä tiedon luokittelumalleja ja eri sovellusalueiden sanastoja kuvauksineen (*ontologiat*). Joitakin semanttisen webin periaatteiden mukaisia www-sivustoja on jo toteutettu. Eräs tällainen on Helsingin tietotekniikan tutkimuslaitoksen ja yliopiston tietojenkäsittelytieteen semanttisen las-

kennan tutkimusryhmän Museo-Suomi -järjestelmä ³. Myös muutamia isoja kaupallisia toimijoita kuten Sun, Adobe ,HP ja IBM on mukana kehitystyössä.

³<http://museosuomi.cs.helsinki.fi/>

10 Yhteenveto

Www-sivut sisältävät paljon maantieteellistä informaatiota, joukossa myös sijaintitietoa (yleensä osoitteita). Erilaisten lähellä olevien kohteiden ja palveluiden hakeminen on näin mahdollista www-hakukoneita hyväksi käyttämällä. Käyttämällä yhtenä hakusanana omaan sijaintiin liittyvää paikannimeä, kasvatetaan sijaintitiedon löytymisen todennäköisyyttä hakukoneen palauttamilta www-sivuilta.

Tavalliset hakukoneet perustuvat hakusanojen käyttöön, eivätkä siis sovellu sellaiseen sijaintitiedon etsimiseen www-sivuilta. Tämän tutkielman yhteydessä toteutettiin koesovellus, joka etsii tavallisen hakukoneen tuottamilta tulossivuilta sijainti- ja paikatietoa. Sovelluksen toinen tavoite oli mahdollistaa tiedonhaku ja tulosten esittäminen pieniresurssisella, mobiililla päätelaitteella.

Toteutuksen pohjana on Google hakukone. Sovellus hakee GPS-laitteelta käyttäjän sijaintikoordinaatit ja muuntaa ne paikkakunnan nimeksi. Nimeä käytetään automaattisesti ohjelman sisällä toimivan Googlen yhtenä hakuehtona. Googlen palauttamilta tulossivuilta etsitään käyttäjän antamiin muihin hakusanoihin liittyvää sijainti- ja paikatietoa. Tulokset esitetään käyttäjälle tiivistetyssä muodossa.

Sovelluksen testaus paljasti toteutuksessa muutamia ongelmia. Yksinkertainen osoitteen etsintäalgoritmi ei löytänyt kaikkia löydettävissä olevia osoitteita. Käytettävyyden parantamiseksi tehdyt rajaukset hakusivujen määriin jättivät relevanttia tietoa hakujen ulkopuolelle. Lukuisat verkon yli otettavat yhteydet ja verkossa olevien resurssien ajonaikainen käyttö altistavat sovelluksen yhteyshäiriöille.

Hakusanojen käyttöön perustuviin www-hakukoneisiin liittyvät tyypilliset ongelmat tulivat korostetusti esiin koesovelluksen yhteydessä. Hakukoneen palauttamien sivujen joukossa on paljon käyttäjän kannalta hyödyttömiä sivuja. Hakusanat kyllä esiintyvät tulossivuilla, mutta eri merkityksessä kuin käyttäjä on ajatellut. Sovelluksessa tietoa haetaan ainoastaan niiltä sivuilta, joilta löydetään sijaintitietoa. Tämä parantaa mahdollisuuksia löytää hyödyllistä tietoa, mutta ei takaa onnistumista.

Koska sovellus on tarkoitettu käytettäväksi mobiililla päätelaitteella, tulisi haun tuloksena saatujen www-sivujen sisältämä tieto esittää tiivistetyssä muodossa. Kun käyttäjä hakee sijaintiaan lähellä olevia palveluja ja kohteita, sovellus muodostaa www-dokumenttien sisällöstä hakutuloslistan. Jokainen hakutulos pyrkii sisältämään haetun

kohteen sijaintitiedot, kohteen kuvauksen sekä sen www-sivun linkin, josta tiedot löytyivät. Johonkin tiettyyn osoitteeseen liittyvän kohteen kuvauksen löytäminen tavalliselta www-sivulta on haastava tehtävä. Koesovellus onnistui testitapauksien yhteydessä tässä kohtuullisesti.

Vaikka hakutuloksen muodostaminen onnistuisi sekä sijaintitiedon että kuvaustiedon osalta, hakutulos saattaa silti olla hyödytön käyttäjälle. Ongelmia aiheuttavat yleensä erilaiset palvelusivustot, joilla saman alueen (esimerkiksi pienen kunnan) kaikki palvelut on koottu samalle sivulle. Sivulta löytyy runsaasti osoitteita ja kuvauksia, joista suurin osa käyttäjän kannalta hyödyttömiä. Löydetyn paikkatiedon hyödyllisyyttä käyttäjälle on vaikea arvioida ohjelmallisesti. Luvussa 8.3 on esitelty muutamia ideoita ongelman ratkaisemiseksi.

Nykymuotoisessa www:ssä hakukoneiden tuottamien tulosten laatua on vaikea parantaa. Www:n sisältämä tietomäärä on valtavan suuri ja tieto täysin järjestäytymätöntä. Tiedonhaun tehostaminen on mahdollista kehittämällä uusia www-sivujen toteutustapoja. Eräs tulevaisuuden mahdollisuus on semanttinen web. Yleistyessään semanttinen web mahdollistaisi tiedonhaun, joka perustuu enemmän merkityksiin kuin yksittäisten merkkijonojen etsimiseen. Www:n hyödyntäminen tietovarastona olisi tehokkaampaa ja luotettavampaa. Esimerkiksi sivulla oleva sijaintitieto voidaan määritellä omaksi resurssikseen, joten sen löytämiseksi ei tarvitse tutkia koko sivun lähdekoodia. Semanttisen webin lausumien avulla selviävät sivuilla olevien tietojen väliset yhteydet ja merkitykset. Tämän ansiosta hakutulosten laatu paranee.

Ongelmistaan huolimatta koesovellus osoitti, että sijaintitietoa käyttämällä nykymuotoistakin www:tä pystytään hyödyntämään liikkuvan ihmisen tietovarastona. Jos käyttäjän kiinnostus kohdistuu ei-kaupalliseen tai ei-fyysiseen kohteeseen, www tarjoaa kaupallisia palveluntarjoajia paremmat mahdollisuudet tiedon löytämiseen. Käyttämällä sopivia osoite/koordinaatti -tietokantoja, sijaintitiedon löytäminen www-sivuilta tehostuu. Kartta-aineiston ja navigointiohjelmiston käytöllä on mahdollista toteuttaa LBS-palvelu, joka pystyy ainakin osittain hyödyntämään www:n valtavaa tietomäärää.

Viitteet

- [1] Berners-Lee, T., Cailliau, R., Luotonen, A., Frystyk Nielsen, H., Secret, A. (1994) The World-Wide Web. *Communications of the ACM* **37** (8), 76-82.
- [2] Dana, P.H.(2005) *Geodetic Datum Overview*. WWW-sivusto, The Geographer's Craft Project, Department of Geography, The University of Colorado. <http://www.colorado.edu/geography/gcraft/notes/datum/datum.html> (28.12.2005).
- [3] Dana, P.H.(2005) *Geodetic Projection Overview*. WWW-sivusto, The Geographer's Craft Project, Department of Geography, The University of Colorado. <http://www.colorado.edu/geography/gcraft/notes/mapproj/mapproj.html> (28.12.2005).
- [4] Defence Mapping Agency. (1984) *Geodesy Of Layman*. WWW-sivusto, http://www.ngs.noaa.gov/PUBS_LIB/Geodesy4Layman/toc.htm (28.12.2005).
- [5] Fonecta Oy. (1984) *Fonecta*. WWW-sivusto, <http://www.fonecta.fi/> (4.1.2006).
- [6] GeoTags. (2005) *Geo Tag Elements*. WWW-sivusto, <http://geotags.com/geo/geotags2.html> (26.12.2005).
- [7] Google Inc. (2005) *Google*. WWW-sivusto, <http://www.google.fi/corporate/tech.html> (20.12.2005).
- [8] Han, J., Kamber, M. (2001) *Data Mining: concepts and techniques*. Morgan Kaufmann, San Francisco.
- [9] Hakala, T. (2005) Valesatelliitit paljastavat sijainnin. *Euroopan tiede ja teknologia* 2(1), 22-24. (Saatavana myös: <http://www.tekes.fi/eu/fin/julkaisut/ett/0105/22.html>, (2.1.2006)
- [10] Hariharan, G., Fränti, P., Mehta, S. (2002) Data mining for personal navigation. *Proceedings of SPIE, Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV* (toim. Dasarathy, B. V.), SPIE, 355-365.

- [11] Heaps, H.S. (1978) *Information Retrieval. Computational and Theoretical Aspects*. Academic Press Inc, New York.
- [12] Herman, I. (2005) *Towards the Semantic Web*. WWW-sivusto, <http://www.w3.org/2004/Talks/0209-Helsinki-IH/> (13.1.2006).
- [13] Kontio, M. (2002) *Mobile Java with J2ME*. Edita Publishing Inc, Finland.
- [14] Maanmittauslaitos. (2005) *Karttapaikka*. WWW-sivusto, <https://www.karttapaikka.fi/karttapaikka/> (24.1.2006).
- [15] Maanmittauslaitos. (2005) *Koordinaatti- ja korkeusjärjestelmät*. WWW-sivusto, <http://www.maanmittauslaitos.fi/Default.asp?id=102> (29.12.2005).
- [16] McCurley, K.S. (2001) Geospatial Mapping and Navigation of the Web. *Proceedings of the 10th international conference on World Wide Web*, ACM Press, 221 - 229.
- [17] Miller, E. (2004) *The Semantic Web*. WWW-sivusto, <http://www.w3.org/2004/Talks/0120-semweb-umich/Overview.htm> (15.1.2006).
- [18] PAM Project Group. (2003) *Personal Navigation Service Architecture: Survey of Background and Standards*. Personal Navigation Programme 2000-2002 (Saatavana myös: http://www.naviverkosto.org/cd/PAM_D6_2_0.pdf, (29.12.2005)
- [19] PHP Documentation Group. (2005) *History of PHP and related projects*. WWW-sivusto, <http://fi.php.net/manual/fi/history.php#history.php> (22.12.2005).
- [20] Rainio, A. (2003) *Paikannus mobiilipalveluissa ja sovelluksissa*. Teknologiakatsaus 143/2003, TEKES (Saatavana myös: http://www.tekes.fi/julkaisut/paikannus_mobiilipalveluissa.pdf, (29.12.2005)
- [21] Schilit, N.B., LaMarca, A., Borriello, G., Griswold, G.W., McDonald, D., Lazowska, E., Balachandran, A., Hong, J., Iverson, V. (2003) Challenge: ubiquitous location-aware computing and the "place lab" initiative. *Proceedings of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, ACM Press, New York, 29 - 35.

- [22] Sonera. (2006) *Missä olet?*. WWW-sivusto, http://www.sonera.fi/artikkeli2/0,3843,1-fi_h-10705_a-250874,00.html (2.3.2006).
- [23] Suomen Posti Oyj. (2006) *Postinumerohaku*. WWW-sivusto, <http://www.verkkoposti.com/e3/postinumeroluettelo> (27.1.2006).
- [24] Tekniikan Sanastokeskus. (2002) *Paikannussanasto*. WWW-sivusto, Tekniikan sanastokeskus. <http://www.tsk.fi/fi/info/paikannussanasto.pdf> (29.12.2005).
- [25] Virrantaus, K., Markkula, J., Garmash, A., Terziyan, V., Veijalainen, J., Katanosov, A., Tirri, H. (2001) Developing GIS-supported location-based services. *Proceedings of the Second International Conference on Web Information Systems Engineering*, IEEE, (2), 66-75.
- [26] Vänskä, I. (2004) Paikkatiedon käyttö web-dokumenteissa. Pro gradu -tutkielma, Tietojenkäsittelytieteen laitos, Joensuun yliopisto.
- [27] World Wide Web Consortium. (2004) *RDF Primer*. WWW-sivusto, <http://www.w3.org/TR/rdf-primer/> (15.1.2006).
- [28] World Wide Web Consortium. (2005) *The World Wide Web Consortium*. WWW-sivusto, <http://www.w3.org/> (22.12.2005).