KANSAINVÄLISTEN PATENTTITIETOJEN ETSIMINEN VOI OLLA SALAPOLIISITYÖTÄ

Samuli Karevaara

Yritysten arvonmuodostuksessa aineettoman pääoman osuus on ollut jatkuvasti kasvussa. Luovan talouden tutkimusyksikkö selvitti Dataamo-hankkeessa muun muassa sitä, mitä kaikkea suomalaiset ovat maailmalla patentoineet. Tässä artikkelissa kerrotaan patenttitietojen etsimisessä kohtaamistamme haasteista.

Patentteja poimimassa

Luovuus yhdistetään usein taiteisiin ja kulttuuriin (Naumanen ym. 2023). Patentoiminen taas tuo monen mieleen vain tekniset laitteet ja suuryritykset, ja tavaramerkit yhdistetään suuriin ja tunnettuihin kuluttajabrändeihin. Todellisuudessa patenttien ja tavaramerkkien kirjo on laaja, ja niiden avulla myös pienet ja keskisuuret yritykset voivat kehittää liiketoimintaansa. Iso osa aineetonta pääomaa on yrityksen ”intellectual property”, kuten patentit, mallisuojat, tekijänoikeudella suojattu aineeton omaisuus tai tavaramerkit (WIPO 2022).

Jotta patenttikenttää ymmärrettäisiin paremmin, Luovan talouden tutkimusyksikössä Dataamo-hankkeessa on louhittu tietoja patenteista sekä Suomessa että ulkomailla. Yhtenä tavoitteena on ymmärtää nykyistä suomalaista innovaatioekosysteemiä ja sen toimintaa paremmin. Tätä tavoitetta edistettiin selvittämällä muun muassa sitä, millaisia keksintöjä suomalaiset ovat maailmalla patentoineet. Kansainvälisten patenttien tarkastelua varten hankittiin lisenssi lens.org-palveluun, josta löytyy yli 147

Karevaara, S. 2024. Kansainvälisten patenttitietojen etsiminen voi olla salapoliisityötä. Teoksessa Rajahonka, M. & Haapaniemi, H. (toim.) Luovia menetelmiä ja älykkäitä ratkaisuja. Digitaalisen talouden vahvuusalajulkaisu 2023. Mikkeli: Kaakkois-Suomen ammattikorkeakoulu, 445–450. https://urn.fi/URN:ISBN:978-952-344-568-0

miljoonaa kansainvälistä patenttitietoa. Hankkeen tavoitteena oli kuitenkin tarkastella suomalaisten patentointivilkkautta. Tätä tarkastelua varten vuonna 2021 Luovan talouden tutkimusyksikkö ja data-analytiikan koulutuslinja alkoivat luoda suomalaista aineetonta pääomaa käsittelevää tietokantaa. Aluksi opiskelijat olivat mukana työssä, ja myöhemmin työtä jatkettiin Otsakorven säätiön tukemassa Dataamo-hankkeessa. Hankkeen aikana valmistui tietokanta nimeltä ”IPR-Suomi”. Se kokoaa tietoa yrityksistä, patenteista, tavaramerkeistä ja mallisuojista vuodesta 2010 lähtien. (Nieminen 2022.) Kansainvälisistä patenteista haluttiin tarkastella lähinnä niitä, jotka voidaan yhdistää kotimaiseen yritykseen.

Patenttia voi hakea joko yksityishenkilö, yritys tai organisaatio. Dataamo-hankkeessa on koottu eri datalähteistä yritysrekisterin lisäksi kattavan datapaketin suomalaisten yritysten ja organisaatioiden hakemista patenteista, tavaramerkeistä sekä mallisuojista. Suomalaisten organisaatioiden tiedot on haettu Patentti- ja rekisterihallituksen (PRH) ylläpitämästä Yritys- ja yhteisötietojärjestelmästä eli YTJ:stä. Patenttitietoja haetaan kansainvälisestä lens.org-palvelusta. Tietoa tavaramerkeistä ja mallioikeuksista haetaan PRH:n Tavaramerkkitietopalvelusta sekä PRH:n Mallioikeustietopalvelusta. Tietoa rahoitetuista TKI-hankkeista noudetaan Euroopan unionin dataportaalista sekä rakennerahastojen julkistamista edellisen ja nykyisen rahoituskauden projektien tietopalveluista.

Eri tietolähteistä kerätty tieto tulisi pystyä linkittämään toisiinsa joillakin muuttujilla. Kaikille yritystietokantaan kerätyille organisaatioiden tiedoille on yhteistä se, että niille on myönnetty suomalainen Y-tunnus. Kansainvälisessä patenttitietokannassa ei kuitenkaan ole suomalaista tai muunkaan maalaista Y-tunnusta, vaan patentinhaltija on nimetty tekstikentän avulla. Tarkasteltaessa suomalaisten maailmalla rekisteröimiä patentteja on kansainväliset patentit ensin yhdistettävä suomalaisesta yritysrekisteristä löytyvään yrityksen nimeen. Yrityksen yhdistäminen nimen perusteella kansainvälisestä patenttihakemuksesta kotimaisessa yritysrekisterissä olevaan yrityksen nimeen tuntuu yksinkertaiselta, kunnes dataa katsoo tarkemmin.

Kaikkien, ainakin kaikkien suomalaisten, tuntema Nokia on vuosien kuluessa patentoinut ahkerasti. Nokia on perustettu vuonna 1865, ja se on historiansa aikana toiminut hyvin monella eri alalla sekä monien eri brändinimien ja liiketoimintayksiköiden alla. Esimerkiksi Nokian tietoliikennetoimintaa suunnitteleva liiketoimintayksikkö on nykyään nimeltään Nokia Networks, mutta aiemmin se tunnettiin nimellä Nokia Solutions and Networks. Kansainvälisestä patenttitietokannasta löytyykin useita

patentteja nimellä ”NOKIA SOLUTIONS AND NETWORKS OY”. Kun nämä patentit yhdistettiin IPR Suomen yritystietokantaan, ilmestyivät Nokian tietoliikenneyksikön patentit tietopankkiin saataville.

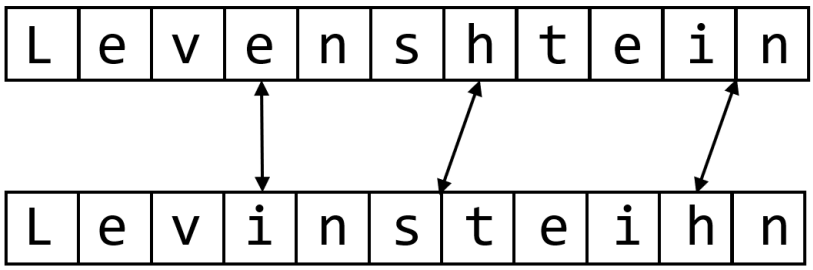
Tietopankin testausjakson jälkeen huomattiin kuitenkin, että Nokia Networksin patentteja puuttui yhä tietokannasta. Tarkemman tutkiskelun jälkeen patentteja löydettiin esimerkiksi nimillä ”NOKIA SOLUTIONS AND NETWORKS” sekä ”NOKIA SOLUTIONS AND NETWORKS HOLDINGS USA INC”. Eli lopun ”OY” puuttui osassa nimistä, ja osassa tilalla oli jonkin toisen maan osakeyhtiötä kuvaava nimen osa, esimerkiksi Saksassa ”GMBH”. Tässä vaiheessa oli selvää, että samanlainen nimeämishaaste koskee myös muita yrityksiä. Samalla huomattiin, että nimen ”AND” oli välillä korvattu &-merkillä ja välillä ei. Tämän osalta tehtiin ratkaisu, jossa niin sanotut hukkasanat (englanniksi ”stop words”) suodatetaan nimestä ensin pois sekä patenttien hakijoiden nimistä että IPR Suomen organisaatioiden nimistä, ja jäljelle jäävää nimen perusosaa käytetään patentin ja patentin hakijan yhdistämiseen.

Nimeämiskäytännöt monimutkaistuvat

Tämän hukkasanojen poistoratkaisun jälkeenkin Nokia Networksin patentteja näytti puuttuvan. Tietokantojen tonkiminen paljasti, että patentteja löytyi myös nimillä ”NOKIA SOLUTIONS AND NETWORKS QY”, ”NOKIA SOLUTIONS AND NETWORKS YO” ja ”NOKIA SOLUTIONS AND NETWORKS 0Y”. Viimeisessä tapauksessa ”0Y” on ”nolla” ”yy”. Kun tietokannasta löytyivät lisäksi nimet ”NOKIA SOLUTIONS AND NETWOKS OY” ja ”NOKIA SOLUTON AND NETWORKS OY” sekä eräänlaisena kirsikkana datakakun päällä ”NOKIA SOLJUSHNZ EHND NETUORKS OJ”, oli selvää, ettei hukkasanaratkaisu tule riittämään. Kyllä, esimerkit ovat aitoja nimiä patenttien omistajista maailmalla.

Kun ongelma monimutkaistuu, usein ratkaisutkin monimutkaistuvat. Hakukoneiden alkutaipaleella ne eivät löytäneet tuloksia, jos hakusanan kirjoitti väärin. Vähän kuin Wordin sanahaku nykyään, jos painat Ctrl + F ja haet sanaa ”finalnd” et löydä sanaa ”finland”. Tämä on hyvä asia silloin, kun tiedät tarkalleen mitä etsit. Entä hakukoneet? Sanahan voi olla väärinkirjoitettuna jollakin verkkosivulla, mutta käyttäjä kuitenkin haluaisi löytää ne sivut, joilla hakusanojen aihetta käsitellään, ei ainoastaan niitä sivuja, joilla hakusanat löytyvät täsmälleen samalla tavalla kirjoitettuna.

Hakukoneet tarvitsivat tavan arvioida sitä, kuinka todennäköisesti sanan kaksi eri tavalla kirjoitettua muotoa yrittävät kuvata samaa asiaa. Tähän voidaan käyttää sumeaa logiikkaa. Siinä tietokoneiden rakastaman joko-tai-logiikan sijasta asiat voivat olla myös osittain jotakin: kuva on 78 % kissa tai sanat ovat 97-prosenttisesti samat. Sanojen vastaavuutta voidaan arvioida esimerkiksi laskemalla sanojen välinen Levenšteinin etäisyys. Se on kirjainmuokkausten pienin mahdollinen määrä, jolla lähdesana voidaan muuttaa kohdesanaksi.



Kuva 1. Esimerkki kahden sanan välisestä Levenšteinin etäisyydestä. (kuva muokattu lähteestä Schraagen & Hoogeboom 2011)

IPR Suomi on rakennettu Power BI -raporttina, mutta datankäsittelyssä käytetään paljon Python-ohjelmointikieltä. Dataamo-hankkeessa laskettiin Python-kielen avulla patenttitietokannasta löytyvien patentinomistajien ja yritystietokannan nimien väliset Levenšteinin etäisyydet ja niiden avulla arvioidut sanojen ”täsmäämisprosentit”. Tämän prosessin aikana huomattiin uusia kotimaisen ja ulkomaisen tietokannan eroja, kuten vanhoista kansainvälisten hiihtokilpailujen tekstityksistä tuttu ”Hämäläinen” on ”Haemaelaeinen”. Jos vain korvaa kirjainparin ”ae” kirjaimella ”ä”, tulee tehneeksi enemmän uusia virheitä kuin korjaa vanhoja.

Menetelmällä laskettuna esimerkiksi ”leikkisät design” on 90,9-prosenttisesti sama kuin ”leikkisaet design”. Ihminen huomaa tästä, että nimet pyrkivät olemaan samat. Varsinkin jos tietää tuon ä → ae -muunnoksen. Kuitenkin noin 91 prosentin match on tässä yhteydessä liian vähän. Jos kaikkia yli 90 prosentin vastaavuuksia pitää samoina niminä, tulee sekaan liian paljon samankaltaisia yritysten nimiä, jotka ovat kuitenkin eri yrityksiä. Dataamo-hankkeessa kokeiltiin myös ChatGPT-tekoälyä täsmäyttämisen apurina, mutta vaikeimpia tapauksia sekään ei tunnistanut samoiksi yrityksiksi.

Jopa ihmisen oli välillä vaikea tunnistaa yritysten väärinkirjoitettuja nimiä. Tunnistaisitko sinä esimerkiksi, mitä yritystä tämä teksti tarkoittaa:

”AJEHJCHK JU INNOVEJSHNS KHEHDKVORTERZ OJ”? Tuo kirjainsoppa toistui hieman eri tavoilla usein, ja tuosta erottui kuvioksi ”aj”, ”ehjch”, ”kju”, ”innovejshns khedkvorterz” ja ”oj”. Tämä taas tunnistettiin loppujen lopuksi englanninkieliseksi ääntämiseksi kirjaimista I, H ja Q sekä ”innovation headquarters oy”. Ja toden totta, IHQ Innovations Headquarters Oy -niminen yritys löytyi IPR Suomen tietokannasta.

Jatkotutkimuksia odotellessa

Koska mikään ratkaisu ei yksinään tuntunut toimivan riittävän hyvin, päädyttiin monitasoiseen hybridiratkaisuun. Yritysten nimistä poistetaan ensin yleisimmät hukkasanat kaikkein ilmiselvimpien eri tavalla kirjoitettujen yritysten nimien löytymiseksi. Sitten sumean täsmäyttämisen logiikalla haetaan yli 99 prosentin todennäköisyydellä samaa yritystä kuvaavat merkkijonot. Lopuksi tarkistetaan tunnettujen tapausten luettelosta aiemmin löydetyt ja ihmisvoimin ilmiselvästi samaa tarkoittavaksi havaitut yritysten nimiparit.

Tämä mukailee hieman myös tekoälyn toimintatapaa. Sillä on tietyt matemaattiset periaatteet, joita se noudattaa. Tämän lisäksi tekoälyä on koulutettu vastauksissaan korostamaan juuri niitä vastauksia, jotka ihmiset kokevat hyviksi vastauksiksi. Väärin kirjoitettuja yritysten nimiä voidaan ”opettaa” järjestelmälle yksinkertaisesti keräämällä ne koneelle valmiiksi lunttilistaksi, josta kurkata, kun tulee tenkkapoo.

Tekoäly on todennäköisesti jatkossa avain tämänkaltaisten salapoliisiongelmien ratkaisemiseen. Kuten tuo IHQ Innovations Headquarters Oy:n tapaus osoittaa, pelkkä tekstianalyysi ei riitä, vaan sanojen ja kirjaimien ääntäminen sekä äänen litterointi takaisin tekstiksi tulee myös ottaa huomioon. ChatGPT:n versioon neljä on lisätty rajapintoja kuvan ja äänen käyttämiseksi syötteenä, joten jatkohankkeen teemaksi ja testattavaksi jää niin sanotun multimodaalisen eli tekstiä, ääntä, kuvaa ja jopa videota hyödyntävän tekoälyn käyttäminen luovasti tämän ongelman ratkaisussa. Ehkä tekoäly voittaa tälläkin saralla pian ihmisen.

LÄHTEET

Naumanen, M., Vainikainen, S. & Valkokari, K. 2023. Tilannekuva luovien alojen ja tapahtuma-alan liiketoiminnasta. Raportti 4/23. Helsinki: VTT.

Nieminen, M. 2022. Eläköön selkeästi visualisoitu data! Xamk NEXT. Kaakkois-Suomen ammattikorkeakoulu. Verkkojulkaisu. Saatavissa: https://next.xamk.fi/uutta-luomassa/elakoon-selkeasti-visualisoitu-data/ [viitattu 2.9.2023].

Schraagen, M. & Hoogeboom, H.J. 2011. Predicting record linkage potential in a family reconstruction graph. Teoksessa Proceedings of 23rd Benelux Conference on Artificial Intelligence (BNAIC 2011), 199-206.

WIPO. 2022. World Intellectual Property Report 2022: The direction of innovation. Geneve, Sveitsi. WIPO.