

Johannes Martikkala

INTERNETIN TIEDONLOUHINTA

Webin datapohjainen analysointi

Kandidaatintyö
Informaatioteknologia ja viestintä
Helmikuu 2020

TIIVISTELMÄ

Johannes Martikkala: Internetin tiedonlouhinta
Kandidaatintyö
Tampereen yliopisto
Tietotekniikka
Helmikuu 2020

Tässä työssä käsitellään internetin tiedonlouhintaa eli webin louhintaa (engl. web mining). Työ käsittelee tiedonlouhinnan yleistä teoriaa, minkä pohjalta muodostetaan kattavaa näkemystä webin louhinnan tutkimusalueen kehittymisestä ja määrittelystä. Webin louhinta jaetaan kolmeen eri osa-alueeseen, joiden käyttötarkoituksia ja eroavaisuuksia tarkastellaan. Työssä käsitellään käytännön sovelluksia eri webin tiedonlouhinnan kategorioista, missä tutkitaan yleisimpiä ja tunnetuimpia webin tiedonlouhinnan sovelluksia. Työ tarkastelee sovellusten toteutustapoja ja käsittelee webin louhinnan etiikkaa.

Työssä todetaan, että webin louhinta on tutkimusalue, jonka pohjalta kehitetyt työkalut ovat antaneet mahdollisuuden käsitellä webin eri osa-alueita tehokkaalla ja monipuolisella tavalla. Tutkimusalue määritellään tiedonlouhintana, jossa data kerätään webistä. Webin louhinta koostuu kolmesta eri osa-alueesta, joita ovat rakenteen, sisällön ja käytön louhinta. Osa-alueiden sovelluksia on monia, joista tunnetuimpia nykypäivänä ovat webin sisäinen hakukone ja palvelun tarjoajien toteuttama kohdennettu sisältö. Webin louhinta on itsessään eettisesti pitävää, jos se toteutetaan datalla, jota on lupa käyttää.

SISÄLLYSLUETTELO

1. JOHDANTO	1
2. WEBIN TIEDONLOUHINNAN TEORIA.....	2
2.1 Tiedonlouhinta	3
2.2 Webin louhinnan historia	4
3. WEBIN LOUHINNAN LUOKITTELU	5
3.1 Web-sisällön louhinta	5
3.2 Web-rakenteen louhinta	7
3.3 Web-käytön louhinta	8
4. WEB LOUHINNAN SOVELLUKSIA	10
4.1 Hakukoneet.....	10
4.2 Kohdennettu sisältö.....	11
5. WEBIN LOUHINNAN EETTISYYS.....	13
5.1 Web-rakenteen ja sisällön louhinnan yksityisyys	14
5.2 Web-käytön louhinnan yksityisyys.....	14
6. JOHTOPÄÄTÖKSET	16
LÄHTEET	17

1. JOHDANTO

WWW (World Wide Web) eli web on globaali tietoverkosto, joka on noussut viimevuosien aikana vakiintuneeksi osaksi kaikkien elämää. Tämä loputtomasti kasvavan määrän dataa sisältävä verkosto avaa uusia mahdollisuuksia sekä kaupallisille että tieteellisille organisaatioille. Sen avulla pystytään tutkimaan miljoonien käyttäjien tuottamaa sisältöä jatkuvasti uusiutuvassa ympäristössä, minkä pohjalta voidaan tehdä ennenäkemättömän kattavaa analyysiä monesta eri aihealueesta. Tämän takia yhä useampi organisaatio on kiinnostunut webin tarjoaman datan hyödyntämisestä osana toimintaansa, mikä on johtanut webin louhinnan löytämiseen omana tutkimusalueenansa.

Työn tarkoituksena on muodostaa kokonaiskuva webin tiedonlouhinnasta eli vastata kysymykseen ”mikä on webin louhinta?”. Työssä käsitellään webin tiedonlouhinnan eri osa-alueet tutkien niiden tiedonhakutapoja, analyysimetodeja ja käyttökohteita internetissä. Työssä käsitellään teoreettisella tasolla eri sovellusvaihtoehtoja ja pohditaan eri tapoja toteuttaa tiedonlouhintaa. Työn tarkoituksena on myös käsitellä eettisiä kysymyksiä tiedonlouhinnan suhteen samalla, kun verrataan eri osa-alueiden käyttöä reaali-maailmassa.

Luvussa 2 perehdytään webin tiedonlouhinnan teoriaan ja historiaan, jonka pohjalta tarkastellaan tiedonlouhinnan määrittelyä. Luvussa 3 määritellään kaikki kolme eri webin louhinnan osa-alueita. Nämä osa-alueet ovat web-sisällön louhinta, web-käytön louhinta ja web-rakenteen louhinta, ja keskitytään niiden sisäisiin määritelmiin ja käyttötarkoituksiin. Luvussa 4 esitellään kaksi tunnettua webin louhinnan sovellusta, jotka ovat hakukone ja käyttäjälle tarjottu kohdennettu sisältö. Luvussa 5 pohdittiin webin louhinnan eettisyyttä eri webin louhinnan osa-alueiden näkökulmasta ja luvussa 6 käsitellään työn johtopäätökset.

2. WEBIN TIEDONLOUHINNAN TEORIA

WWW eli web on nykyaikainen väline jakaa informaatiota, mikä sisältää suuria määriä tietoa eri aiheista eri muodoissa. Webin suuren koon ja heterogeenisen sisällön perusteella on sen käsitteleminen haastavaa. Näiden haasteiden myötä on vaikea kerätä tarkoituksenmukaista informaatiota, kohdennettua informaatiota tai kehittää uutta tietämystä halutusta aihealueesta. Monien eri teknologioiden, kuten tietokantojen, hyödyntäminen osana web-aplikaatiota tai yksinkertaiset tiedonhaut ovat yrittäneet ratkaista näitä haasteita. Mutta yksinään mikään ratkaisu ei ole ollut yksiselitteinen vastaus webin asettamiin haasteisiin. [1]

Vastaavasti webin loputtomasti kasvava datamäärä, joka on jokaisen saatavilla ilmaiseksi, on tarjonnut loputtoman potentiaalin dataan perustuvaan analyysiin. Tätä analyysiä voidaan käyttää osana kaupallista liiketoimintaa ja sen analyysiä tai ei-kaupallisesti osana tieteellistä tutkimusta. Tämän myötä webin louhinta tutkimusalueena on kerännyt laajan huomion sekä tieteellisissä tutkimuksissa että kaupallisten tahojen toteuttamissa tutkimuksissa, mikä on johtanut webin tiedonlouhinnan eli webin louhinnan muodostumiseen omaksi tutkimusalueekseen. [2]

Tiedonlouhinnalla tarkoitetaan monivaiheista prosessia, jonka tarkoituksena on kerätä tarvittava data ja tuottaa sille tarvittavaa käsittelyä, kuten sen mallintamista ja analyysiä, minkä jälkeen lopulta saadaan tuotettua haluttua informaatiota käsittelyssä olleesta datasta. Prosessin vaiheet ovat tyypillisesti datan hankkiminen, datan esiprosessoiminen, datan mallintaminen, datan tulkitseminen ja tulosten arviointi. Prosessin datan laadulla on merkitystä eikä sen lähteellä, minkä takia webistä hankittu data on pätevä kohde tiedonlouhinnan prosessille. [3]

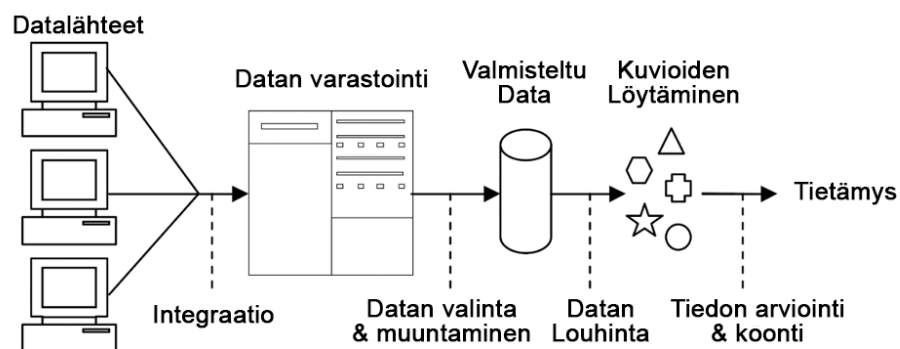
Webin tiedonlouhinta on tiedonlouhinnan laajennettu versio, jossa hyödynnetään tiedonlouhinnan ratkaisutapoja internetin analysointiin. Pieniä eroavaisuuksia näiden kahden tutkimusalueen välillä voi esiintyä. Merkittävin eroavaisuus tutkimusalueiden välillä on tiedon tallentaminen, joka tapahtuu tiedonlouhinnassa tietokannoissa ja webin louhinnassa palvelimilla. Koska eroavaisuudet ovat pieniä, voidaan webin louhinnan tutkimus-alueetta pitää suoraan tiedonlouhinnan jatkokehitysalueena. [2]

2.1 Tiedonlouhinta

Tiedonlouhinnan tutkimusalue käsittelee eri tekniikoita, joilla kerätään tietoa datasta. Tiedonlouhinta ei käsittele datan keräämistä, vaan keskittyy sen käsittelemiseen. Tämä prosessi on moniosainen, ja yleisesti se sisältää seuraavat vaiheet:

1. Datalähteiden siivoaminen. Poistetaan virheelliset ja epä johdonmukaiset data-lähteet.
2. Datan integraatio. Yhdistetään mahdollisesti monen eri lähteen datat keskenään.
3. Datan valinta. Valitaan analyysille oleelliset datakohteet.
4. Datan muuntaminen. Muunnetaan data analyysille sopivaksi erilaisilla operaatioilla.
5. Datan louhinta. Oleellinen ja tapauskohtainen prosessi, jossa havainnoidaan datasta löytyviä malleja.
6. Mallien arviointi. Tarkastellaan saatuja tuloksia ja tehdään sen perusteella arviointia analyysistä.
7. Tiedon esittely. Esitellään visualisoinnin ja muiden esitystapojen avulla tehtyjä tuloksia.

Kohdat 1–4 ovat datan esivalmisteluvaihteita, joissa dataa valmistellaan sopivaksi itse louhintaa varten. Louhinnan jälkeiset vaiheet on tarkoitettu siihen, että tarkastellaan, onko louhinta toteuttanut tarkoituksensa, ja louhinnan tulokset esitetään helposti ymmärrettävässä muodossa. Prosessin rakenteen perusteella voidaan todeta, että itse datan louhinta on tiedonlouhinnan oleellisin kohta [4]. Tiedonlouhinnan prosessikaavio on esitetty alla olevassa kuvassa 1.



Kuva 1. Tiedonlouhinnan prosessikaavio [5].

On kuitenkin huomioitava, että datan louhinta ei yksinään määrittele tiedonlouhinnan onnistumista. Sitä edeltävät prosessit mahdollistavat halutun louhinnan toteuttamisen ja

sen jälkeiset vaiheet vahvistavat prosessin koko työn. Varsinkin alkupään vaiheet prosessissa ovat kriittisiä, koska vääränlaiset datalähteet, puutteelliset datasetit tai väärin valitut datat mitätöivät koko muun prosessin. [5]

2.2 Webin louhinnan historia

Webin louhintaa on määritelty kahdella eri tapaa. Webin louhinta voidaan määritellä joko prosessikeskeisesti tai datakeskeisesti [6]. Prosessikeskeisessä määrittelyssä webin louhintaa tarkastellaan prosessina. Prosessissa suoritetaan määrättyssä järjestyksessä tehtäviä. Prosessia kutsutaan webin louhinnaksi. Prosessin vaiheita ei ole määritelty yhtenäisesti, mutta vaiheet on tyypillisesti jaettu kolmeen osaan: resurssin löytämiseen, informaation keräämiseen ja sen käsittelyyn [7].

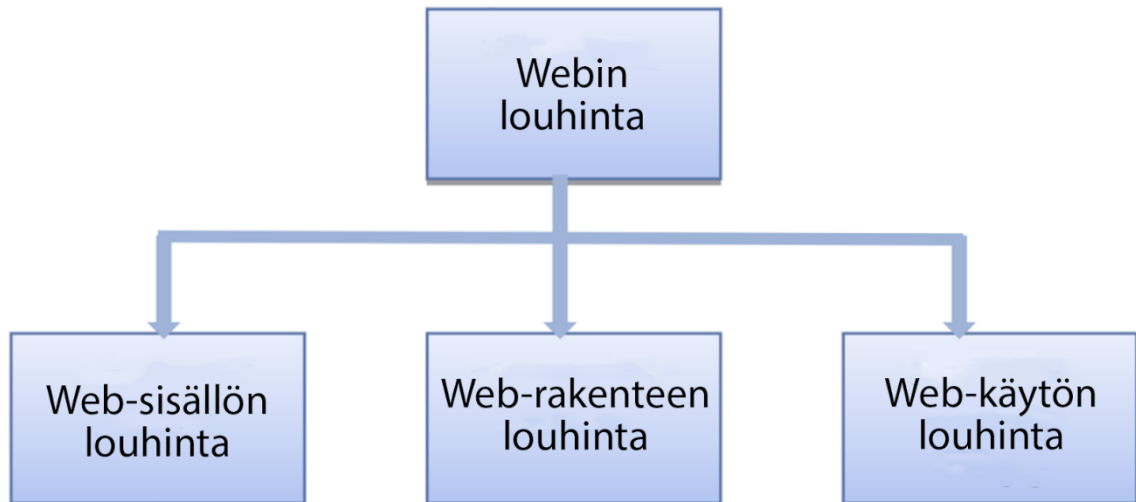
Datakeskeisessä määrittelyssä webin louhintaa määritellään niin, että se on tiedonlouhintaa, jossa käytetään webistä kerättyä dataa [8]. Data voi olla laadultaan vaihtelevaa, mutta sen lähde on aina internetistä. Tämä määrittely on paljon yleisempi tapa määritellä webin louhinta, mikä ilmenee aiheesta tehdyistä tutkimuksista, joissa webin louhinta määritellään tiedonlouhintana, jossa data saadaan webistä [6].

Tällöin voidaan todeta, että webin louhinnan tärkein komponentti on ollut kehityksessä jo pitkään, vaikka webin louhinta on itsessään uudempi tutkimusalue. Kuten muut webin tutkimusalueetkin, on tietotekniikan kehittymisen myötä tiedonlouhinnan tutkimusalue keskittynyt käsittelemään jatkuvasti suurempia määriä dataa löytääkseen mahdollisimman optimaalisia tapoja tuottaa kattavaa analyysiä. [2]

Sen lisäksi, että webin louhinta on pitkälti pohjautunut tiedonlouhintaan, on huomioitavaa, että webin louhinnan tutkimusalue on myös hyötynyt tekstin louhinnan tutkimusalueesta suuresti. Tekstin louhinta on tekstin analysointiin erikoistunut tutkimusalue, jossa keskitytään eri tapoihin analysoida tekstiä. Tämän tutkimusalueen saavutuksia hyödynnetään suuresti, kun analysoidaan webin tarjoamaa sisältöä, koska suurta määrää tekstiä käsitellessä on oleellista käsitellä teksti tehokkaasti ja niin, että oleellimmat asiat on saatu hahmoteltua tekstistä. Tämän avulla pystytään webin sisällöstä muodostamaan oleellisia koosteita mahdollistaen itse webin analysoinnin, mikä on hyödyllistä, koska todella suuri osa webin sisältämästä datasta esiintyy tekstin muodossa. [9]

3. WEBIN LOUHINNAN LUOKITTELU

Webin tiedonlouhinta voidaan jakaa kolmeen alakategoriaan perustuen siihen, mitä osaa webistä halutaan louhia. Webistä voidaan louhia tietoa sen sisällöstä, rakenteesta ja sen käytöstä. Nämä kolme kategoriaa voidaan vielä sisäisesti jaotella pienemmiksi kategori-
oiksi perustuen niiden tiedon laatuun [10]. Tämä jaottelu on esitelty kuvassa 2.



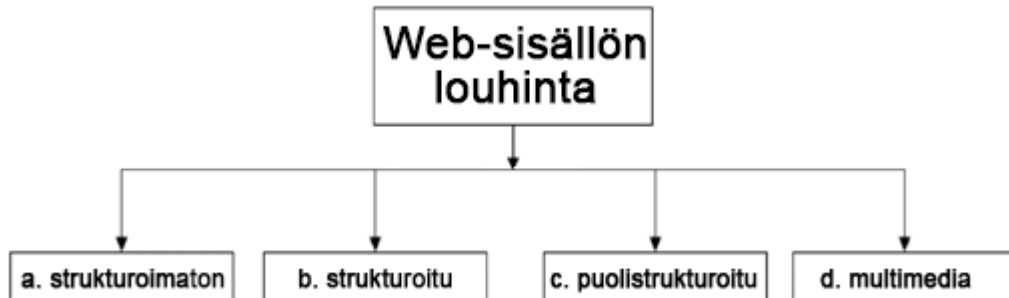
Kuva 2. Webin tiedonlouhinnan jaottelu [11].

Louhinnassa voidaan yhdistellä eri kategorian asioita keräten tietoa eri osista yhtäaika-
sesti. Näillä eri osa-alueilla on selkeästi eri käyttötarkoituksia, mutta niiden yhdistämisen
myötä voidaan mahdollistaa kattavamman analyysin muodostaminen.

3.1 Web-sisällön louhinta

Webin yksi tärkeimmistä louhinnan kategorioista on sisällön louhinta. Sisällön louhimi-
sessä käsitellään webin tarjoamaa primääriä sisältöä, eli nettisivujen sisältöä. Koska we-
bin tarjoama data ei ole mitenkään säänneltyä, on sen tarjoama sisältö hyvin haasteelli-
nen kohde tiedonlouhinnalle. Webin tarjoama data vaihtelee struktuuriltaan, kontekstil-
taan, tyypiltään ja lähteeltään suuresti, mutta sen lisäksi se sisältää toistuvaa tietoa ja on
jatkuvasti muuttuvaa. Tämän takia on web-sisällön louhinnassa kiinnitettävä huomiota
hankittavan datan luonteeseen yhtä paljon kuin sen lähteiseen saavuttaakseen haluttua
analyysiä kohteesta. [12]

Web-sisällön louhinta voidaan jakaa neljään eri alakategoriaan louhinnan kohteen olevan sisällön perusteella. Nämä neljä kategoriala ovat strukturoitumaton, strukturoitu, puolistrukturoitu ja multimedia. Nämä kategoriat on esitetty alla olevassa kuvassa 3.



Kuva 3. Webin sisällön louhinnan kategoriat [10].

Strukturoimattoman sisällön louhinnan yleisin tarkoitus on tuottaa uutta tietoa. Louhinnassa oleva strukturoimaton data on usein tekstiä, jota pystytään käsittelemään ohjelmistotasolla, vaikka teksti onkin strukturoimatonta. Usein strukturoimattoman datan käsittelyn yhteydessä käsitellään relevantteja avainsanoja ja lauseita, jotta löydetään analyysin kohteena olevan aihealueen sisältöä. Tarjolla olevasta datasta ja tarkoituksesta riippuen strukturoimattomasta datasta voidaan hakea avainsanojen perusteella uutta tietoa, seurata tiettyjen aihealueiden esiintymistä webissä, luoda tiivistyksiä eri aihealueista, kategorioida tietyn aihealueiden sisältöjä, ryhmittää aihealueiden esiintymiset keskenään tai tuottaa visualisointia aihealueen esiintymisestä. [11]

Strukturoidulla datalla tarkoitetaan tässä tarkoituksessa dataa, joka on ennalta määriteltävissä muodossa tai dataa, jonka esiintymistä halutaan selvittää. Ennalta määriteltävä data on yleensä nettisivujen tuottamaa dataa toisista nettisivuista. Vastaavasti dataa, jonka esiintymistä halutaan selvittää, tarkoittaa dataa, joka kertoo, sisältyykö etsitty hakusana tai lause milläkin nettisivulla. Tällöin puhutaan etsintäperusteisesta tiedosta eli tiedosta, joka kerättiin tutkimalla, että mistä nettisivuista haettu data löytyy. Hakukoneet toteuttavat siis strukturoidun tiedonlouhinta etsiessään, esiintyykö käyttäjän syöttämä hakulause milläkin nettisivulla. [11]

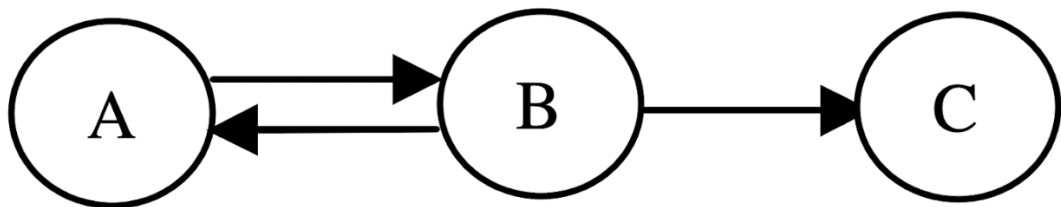
Puolistrukturoidussa tiedonlouhinnassa yhdistetään usein näiden kahden edellisen kategorian muotoja ja tällöin analyysin kohteena on usein joko strukturoitujen nettisivujen strukturoimaton sisältö tai jokin sen variaatioista. Tällöin analyysin kohteena voi olla jokin edellisten kohtien välimallista datan esiintymisen tutkimisesta sen sisällön koontiin. [11]

Multimedian louhinta poikkeaa muun sisällön louhinnasta tyypiltään. Multimediallinen sisältö on suhteellisesti poikkeavaa kuin webin tarjoama tekstimuotoinen data, mikä asettaa haasteita sen analyysille. Multimedian louhinta on siis pitkälti kiinni sen yhteydessä käytetystä analyysityökalusta ja analyysin tarkoituksesta. Näitä ovat esimerkiksi tiettyjen kuvioiden löytämistä multimedialta, kuten tietyn kappaleiden olemassaolon etsimistä tai videonpätkien löytämistä webistä. Vastaavasti multimedian louhinnalla voidaan selvittää erilaisia kuvioita ja niiden esiintymistä eri nettisivuilla. [11]

3.2 Web-rakenteen louhinta

Web-rakenteen louhinnan tarkoituksena on tuottaa analyysiä nettisivujen keskinäisistä suhteista. Tarkastelun kohteena on nettisivujen hyperlinkit, jotka ohjaavat toisiin nettisivuihin tai muihin internetin lähteisiin. Näiden linkkien pohjalta voidaan muodostaa analyysiä nettisivujen välisistä rakenteista, kun pystytään selvittämään, että mistä nettisivusta on mahdollista päästä millekin nettisivulle. [13]

Rakenteen louhinnan perusteella webistä voidaan muodostaa graafeja, joissa solmukohdat ovat nettisivuja ja kaaret ovat hyperlinkkejä nettisivujen välillä. Esimerkki tällaisesta graafista on esitelty kuvassa 4.



Kuva 4. Esimerkki nettisivu-kaaviosta [14].

Kaaret voivat olla joko yksisuuntaisia tai kaksisuuntaisia riippuen siitä, sisältävätkö nettisivut keskenään hyperlinkkejä toisiinsa, vai onko vain toisella nettisivulla hyperlinkki toiseen nettisivuun. Itse solmut, eli nettisivut, vaihtelevat täysin rakenteeltaan, koska yksi nettisivu voi sisältää monta hyperlinkkiä moneen eri nettisivuun tai vastaavasti ei yhteenkään. Vastaavasti myös yksi nettisivu voi olla linkitetty monella sivulla tai ei yhdelläkään sivulla. [15]

Rakenteen louhimisessa voidaan tutkia webin yleistä rakennetta, mutta yleisempää on tuottaa analyysiä yhden nettisivun kontekstista. Tällöin analyysissä on kokonaisen graafin muodostamisen sijaan tärkeää huomioida hyperlinkkien lähteet. Rakenteen louhimisen avulla voidaan yhdistää eri konteksteja välttämällä samalla ylimääräistä toistoa. Yksi

nettisivu voi ilmetä monen nettisivujen hyperlinkeissä, mikä johtaa nettisivujen ylimääräiseen toistoon, jos analyysissä ei otettaisi huomioon nettisivuun johtanutta rakennetta. Mutta koska voimme pitää erillistä kirjaa lähteistä, voimme käsitellä tarkasteltavaa nettisivua vain yhtenä solmuna ja muodostaa monia kaaria siihen, milloin graafista tulee helpommin analysoitava ja analyysistä kattavampaa. [15]

Toisena lähteiden tutkiminen mahdollistaa kattavan kontekstianalyysin tuottamisen nettisivulle. Lähteet ovat keskenään varsin eriäviä, koska nettisivut voivat olla usein eri tekijöiden tekemiä ja sen myötä sisältää erilaista sisältöä. Kun voimme kerätä mahdollisimman monen eri lähteen sisällön, ei kontekstianalyysi riipu merkittävästi yhdestä lähteestä vaan saa kattavamman pohjan hyperlinkin kontekstianalyysille. Kolmantena lähteiden tutkimisella voidaan tutkia, minkä tyyppisillä nettisivuilla tutkittu nettisivu esiintyy. Tämän lisäksi voidaan myös erotella nettisivut, jotka ei tarjoa mitään kontekstia luettavassa muodossa, vaan ovat esimerkiksi hakukoneita, joiden relaatio nettisivuihin ei tarjoa suurta informaatiota nettisivun kontekstista. Rakenteen louhimisen tuottaman tiedon avulla voimme myös eliminoida irrelevantin tai harhaanjohtavan tekstin vaikutuksen analyysissä. Monet nettisivut voivat sisältää täysin irrelevanttia tekstiä hyperlinkkiin nähden, mutta niiden vaikutus pienenee, kun kontekstianalyysissä yhden nettisivun irrelevantti teksti toistuu vain kerran. Tämän lisäksi voimme myös erotella nettisivut keskenään sen sisällön kielen perusteella tuottaen syväsuuntaisempaa analyysiä siitä, että minkä kielisillä nettisivuilla tarkasteltu nettisivu esiintyy. [15]

3.3 Web-käytön louhinta

Webin käytön louhinta on prosessi, jossa tarkastellaan dataa nettisivujen käytöstä. Dataa voidaan kerätä palvelinpuolelta, välityspalvelimen puolelta ja suoraan tai kierteisesti käyttäjältä. Näiden datojen saatavuus on tapauskohtaista. Palvelun tarjoajat voivat päättää, mitä dataa tallentuu palvelun käyttäjästä palvelun palvelimelle. Vastaavasti ulkopuolisten palveluiden tuottama datan saatavuus ja laatu vaihtelevat suuresti. Usein dataa kerätään kaikista mahdollisista lähteistä tavoittaakseen kustakin käyttäjästä mahdollisimman kattavan analyysin. Data tässä louhinnassa on todella monipuolista, koska selaimet ja nettisivut tallentavat vaihtelevasti käyttäjistä dataa riippuen käyttäjästä, käyttäjän käyttöliittymän asetuksista, käytettävästä selaimesta ja historiatiedoista. [16]

Web-käytön louhinta kerätty data voi olla joko kohdennettua yksilöön tai ei kohdennettua yksilöön. Kohdennetussa datassa usein tutkitaan käyttäjän lokitiedostoja, evästeitä ja yksityiskohtaista käyttäytymistä, ja ne silloin liitetään aina yksittäiseen käyttäjään esimerkiksi käyttäjätunnuksen tai IP-osoitteen perusteella. Vastaavasti, kun louhitaan anonyy-

misti dataa, kerätään samoja datapisteitä ilman, että kohdennetaan data yksittäisiin käyttäjiin. Yksilöön kohdennetulla datalla voidaan tutkia yksilöitä ja tuottaa yksilöllistä analyysiä. Vastaavasti ei kohdennettua louhintaa toteutetaan tutkiakseen kaikkia käyttäjiä ja siinä esiintyviä malleja. On myös hyvin yleistä, että analyysissä tutkitaan sekä kohdennettua ja ei kohdennettua dataa, milloin voidaan tuottaa sekä kohdennettua analyysiä että hahmottaa kokonaisuuksia käyttäjäkunnassa. [17]

Web-käytön louhinnan pohjalta voidaan tuottaa analyysejä siitä, että miten käyttäjät käyttäytyvät nettisivuilla. Sen lisäksi voidaan esimerkiksi hahmotella selaustietojen korreloimista käyttäjän tapoihin, ryhmitellä erilaisia käyttäytymistapoja tai tutkia eri eväsetietojen korrelointeja eri käyttäytymismalleihin. Tämän analyysin pohjalta voidaan esimerkiksi ennakoida käyttäjän tarpeita tarjoten tälle häntä todennäköisesti kiinnostavaa sisältöä, analysoida itse palvelun toiminnasta, optimoida sivuston toimintaa kullekin käyttäjätyyppille tai palvelun toiminnan optimointiin perustuen käyttäjän mieltymyksiin. [18]

4. WEB LOUHINNAN SOVELLUKSIA

Webin kasvun myötä on myös webin louhinta saanut enemmän huomiota sekä tutkimusalueena että kaupallisten organisaatioiden apuvälineenä [1]. Tämän myötä on kehittynyt erilaisia sovelluksia, joiden avulla pystytään tuottamaan erilaisia palveluja ja analyysejä perustuen tähän webin louhinnan tuottamaan analyysiin. Nämä sovellukset voivat olla yksittäisiä kokonaisuuksia tai olla osana suurempaa sovellusta.

4.1 Hakukoneet

Yleisimpiä webin louhinnan sovellukset ovat hakukoneet. Hakukoneiden tarkoituksena vastaanottaa käyttäjältä hakusanoja ja etsiä sen perusteella mahdollisimman relevantteja nettisivuja. Koska hakukoneet ovat yleensä suojeltua omaisuutta, on niiden palvelujen sovellusarkkitehtuuri usein salattua tietoa. Kuitenkin voidaan yleisesti olettaa, että hakukoneet käyttävät usein jokaista webin louhinnan eri osa-aluetta yhdessä. Rakenteen louhinnan avulla pystytään tutkimaan sisältääkö nettisivu annettua hakusanaa missään muodossa. Analyysin nopeuttamiseksi tuotetaan usein nettisivun sisällön louhinnan yhteydessä metatietokooste nettisivusta sisältäen kaikki oleelliset avainsanat, jolloin nettisivun sisältöä ei tarvitse jokaisen tiedustelun yhteydessä uudelleen analysoida. Käytön louhinnan avulla hakukoneet voivat hahmottaa, mitkä nettisivut ovat suosituimpia kuin toiset. Tätä analyysiä voidaan tuottaa hakukoneen omien käyttäjien perusteella, milloin voidaan helposti tuottaa informaatiota siitä, mitkä hakukoneet ovat käyttäjien suosimia milläkin hakusanoilla. Ja viimeisenä voidaan tuottaa analyysiä siitä, mitkä nettisivut ovat relaatioissa muihin nettisivuihin ja kuinka paljon relaatioita nettisivu omistaa, milloin voidaan hahmotella nettisivun suosiota. [6]

Esimerkki hakukoneen implementaatiosta on Googlen tunnettu hakukone [19]. Googlen hakukone on indeksoinut biljoonia nettisivuja palvelimelleen mahdollistaen mahdollisimman nopean ja kattavan haun. Googlen hakukone hyödyntää kaikkia eri webin louhinnan osa-alueita, mutta se tuli alun perin tunnetuksi rakenteen louhimisen hyödyntämisestä louhinnan yhteydessä. Googlen hakukone käyttää kehittyntä tapaa järjestää keskenään sivustot prioriteettijärjestykseen PageRank-nimisen algoritmin perusteella, joka määrittelee sivuston tärkeyden. Tämän lisäksi Googlen hakukone hyödyntää sivustojen käsittelyn yhteydessä tuotettua rakenteellista analyysiä itse haun tarkentamiseksi. Tämän lisäksi google myös hyödyntää haussa käyttäjiensä hakuja hahmottaakseen, että mitkä hakujen tulokset ovat toisia suosituimpia. Tämän lisäksi google pitää kirjaa suosituimmista hakusanoista muodostaakseen kuvan siitä, että mitä tietoa netistä kannattaa

louhia. Tätä kutsutaan tekstin ankkuroinniksi (engl. Anchor Text), ja siinä nettisivuston sisällöstä muodostetaan koosteita, että mitä nettisivusto sisältää. Toteuttaakseen tämän, käytetään rakennelouhinnan kontekstianalyysin lisäksi rakenteen louhimista, jonka myötä Google pystynyt tallentamaan palvelimilleen kaiken oleellisen tiedon tehokasta indeksointia varten. [6]

4.2 Kohdennettu sisältö

Webin louhinnan myötä on mahdollistunut myös webin käyttäjien analysointi, mikä on johtanut kohdennetun sisällön tarjoamiseen webin käyttäjälle. Tämä käyttäjien antamaan yleiseen tai yksilölliseen dataan pohjautunut analyysi on kiinnostanut monia eri organisaatioita, koska se mahdollistaa entistä paremmin käyttäjien ohjaamisen ja niille tarjoaman sisällön määrittelemisen. Kohdennetun sisällön prosessi muodostuu viidestä moduulista: käyttäjän profiloitumisesta, web-käytön louhinta, tiedon hankinta, sisällön hallinta ja web-sisällön julkaiseminen. Näistä web louhintaan liittyvät osat ovat kolme ensimmäistä kohtaa, joissa hankintaan tietoa yksilöistä sekä kaikista käyttäjätavoista, minkä pohjalta voidaan hallita sisältöä ja julkaista sitä käyttäjille. [18]

Kriittinen osa henkilökohtaisesti kohdennettua sisältöä on käyttäjän profiloituminen ja web-käytön louhinta. Tätä prosessia web-palvelut tekevät joko implisiittisesti tai eksplisiittisesti. Implisiittisesti käyttäjästä voidaan kerätä tietoa esimerkiksi käsittelemällä käyttäjän jättämiä verkkoloki-tietoja, käyttäjätietoja tai evästeitä. Eksplisiittisesti tietoa voidaan kerätä erinäisten kyselyiden avulla tai kerätä palautetta tarjotusta sisällöstä. Tässä osassa prosessia kerätään kaikki oleellinen tieto käyttäjästä. Koska palvelun tarjoaja vastaa itse palvelusta, on mahdollista, että käyttäjästä kerätään kaikki palvelussa tapahtuva tieto analyysiä varten. [18]

Tämän tiedon avulla voidaan tuottaa kattavaa analyysiä eri asioista. Voidaan hahmotella esimerkiksi näiden käyttäjistä saatujen tietojen avulla eri käyttäytymiskuvia, erilaisia assosiaatioita eri käyttäjätapojen välillä tai luokitella käyttäjiä niiden profiloitumisen perusteella. Analyysi on kuitenkin aina tapauskohtainen ja lopulta perustuu pitkälti siihen, että minkälaista tietoa halutaan tuottaa käyttäjästä. Kuitenkin on huomioitavaa, että koska käyttäjästä voidaan kerätä tietoa sekä oman palveluiden kautta että ulkoisten palveluiden avulla joko suoraan palvelulta tai käyttäjältä, on mahdollista, että analyysi pohjautuu pitkälti muiden palveluiden tuottamaan dataan. Tällöin voidaan tuottaa käyttäjän ja palvelun interaktion ulkopuolelta perustuvaa analyysiä, milloin kohdennettua sisältöä voidaan tarjota jo ensimmäisellä käyttökerralla. Tällaisesta datasta esimerkkejä voi olla

esimerkiksi käyttäjän kansallisuuden antamat paikka- tai selaustiedot, tai suoraan toiselta palvelulta saadut datat. [20]

Tämän analyysin pohjalta voidaan käyttäjälle tarjota kohdennettua sisältöä. Palvelun tarjoajat voivat tuottamansa analyysin pohjalta tarjota käyttäjälleen todennäköisesti optimaalisinta nettisivun sisältöä ja rakennetta perustuen siihen, että mitä louhinnan perusteella käyttäjän profiili voisi palvelulta haluta. Analyysiä tuotetaan jatkuvasti, kun käyttäjä käyttää palvelua, muokaten kohdennetun sisällön tarjontaa. On kuitenkin huomioitavaa, että palvelun laadun mukaan, voi kohdennetun sisällön tarjoaminen tarkoittaa eri asioita. Palvelu voi esimerkiksi tavoitella mahdollisimman vaivatonta käyttäjäkokemusta tai vaihtoehtoisesti tarjota käyttäjälleen mahdollisimman paljon potentiaalisesti kiinnostavia viirikkeitä, milloin sen pohjalle teetetty louhinta vaihtelee laadultaan. [18]

5. WEBIN LOUHINNAN EETTISYYS

Webin louhinta on osoittanut, että webin tarjoamaa dataa voidaan käyttää hyödyllisesti ja se mahdollistaa monien palvelujen ja sovelluksien tuottamista. Käyttäjät voivat käyttää webin palveluita ja hyötyä osin siitä, että heidän datansa mukaan voidaan esimerkiksi tarjota heille relevantimpaa sisältöä ja mahdollistaa optimoidumpien palveluiden tarjoamista. Tämän saavuttamiseksi käyttäjät voivat antaa itsestään suhteellisen persoonallista informaatiota, jos vain saatu vastine on heidän mielestään informaation arvoista ja käytettävä web palvelu on tarpeeksi vakuuttava sen suhteen, että se ei väärinkäytä annettuja informaatioita. [6]

Tämän takia webin louhinnan tärkein eettinen kysymys on se, että käyttäjät eivät voi hallita sitä, että miten hänen dataansa voidaan todellisuudessa käyttää. Webin käyttäjät voivat tuottaa paljon dataa joko käytettyihin palveluihin tai esimerkiksi selaimeen, mutta heillä ei ole mitään tapaa hallinnoida sitä, että kuka voisi hänen dataansa päästä käsiksi tai missä kaikkialla sitä on. Tämä mahdollistaa sen, että käyttäjien dataa voidaan kerätä näkymättömästi ilman heidän suostumustaan, minkä myötä sitä voidaan käyttää ilman käyttäjän antamaa lupaa. Tämän lisäksi on myös mahdollista, että käyttäjän syöttämät tiedot päätyvätkin toiseen käyttöön kuin palvelu, johon tiedot oli syötetty, oli käyttäjälle luvannut. [21]

Käyttäjien yksityisyys on yleinen ongelma webissä. Käyttäjän yksityisyydellä tarkoitetaan yksilön oikeutta hallinnoimaan informaatiota käyttäjästä. Yksityisyyden loukkaamisella tarkoitetaan sitä, kun käyttäjään liitettävää informaatiota kerätään, käytetään tai levitetään ilman, että käyttäjä on tietoinen tästä tai on antanut lupaa siihen. Webin louhinnan yksityisyysongelmat usein ovat liitännäisiä tähän ongelmaan. Louhinnassa kerätty data voi loukata jonkun käyttäjän yksityisyyttä. Louhinnassa käsitelty data voi olla joko yksilöllistä ja johdettavissa suoraan tiettyyn käyttäjäprofiiliin, tai olla luonteeltaan anonyymiä. Kuitenkin on huomioitavaa, että vaikka data olisikin tehty anonyymiksi, voi se silti rikkoa käyttäjän yksityisyyttä, koska vaikka se ei voitaisi suoraan liittää kehenkään käyttäjään, voitaisiin silti datasta löytää hänelle henkilökohtaista tietoa johtaen käyttäjän yksityisyyden rikkomiseen. Ja vaikka datasta ei löytyisi yksilöllisiä tietoja, voisi käyttäjä silti tuntea, että hänen yksityisyytään on rikottu, koska dataa on saatettu käyttää tavalla, johon käyttäjä ei ole antanut lupaa. [21]

5.1 Web-rakenteen ja sisällön louhinnan yksityisyys

Poisluettuna palvelut, jotka piilottavat rakennettaan ja sisältöään muilta käyttäjiltä, on webissä oleva sisältö ja rakenne julkista tietoa. Web sisältää paljon dataa ja nettisivuja, jotka ovat kontekstissaan juuri käyttäjän toivomia ja luomia. Mutta koska louhinta erottaa usein sekä nettisivujen sisällön ja rakenteen alkuperäisestä kontekstistaan, voi tällöin käytettävä informaatio joutua väärinkäytetyksi tai jaettavaksi väärään kontekstiin, mikä johtaisi alkuperäisen käyttäjän yksityisyyden rikkomiseen. Voidaan todeta, että käyttäjän yksityisyyttä ei rikota silloin, jos käyttäjä on alun perin asettanut tiedon julkiseksi ja sitä ei esitellä väärin sen kontekstia. Vasta-argumentti tähän on se, että käyttäjällä on oikeus esittää sisältönsä vain siinä kontekstissa, missä hän alkuperäisesti julkaisi sisältönsä. Sisällön ottaminen pois kontekstista ja liittäminen johonkin sisältöön kuulumattomaan kontekstiin voi johtaa yksilöä loukkaavaan tilanteeseen. Yksiselitteistä ohjesääntöä hyviin tapoihin webin louhinnassa ei ole tehty eikä mikään laji rajaa suoranaisesti sisällön käyttöä. Kuitenkin yleinen käytäntö on, että muiden sisältöä esitellessä on oleellista tarjota alkuperäinen konteksti, koska muuten käyttäjän tiedot voivat tulla väärinkäytetyksi. [21]

Yksityisyyden eettiseen ongelmaan ei ole yksiselitteistä vastausta. Web-sisällön ja rakenteen louhiminen on oikein toteutettuna eettisesti pätevää, mutta voi potentiaalisesti rikkoa käyttäjän yksityisyyttä. Koska web on kaikille avoin ja pitkälti säännöstelemätön ympäristö, on todettavaa, että yksityisyyden suojaamiseksi on hyvin rajatusti vaihtoehtoja. Laki voi antaa perustan, jonka pohjalta voidaan käsitellä rikkomuksia, jos ne havaitaan. Rikkomusten havaitseminen voi olla vaikeaa rikkomuksen mukaan. Väärin käytetyn sisällön havaitseminen on mahdollista esimerkiksi webin louhimisen avulla. Mutta esimerkiksi väärin käytetyt käyttäjäprofiilit, jotka ovat tehty louhimalla käyttäjän tekemää sisältöä, on lähes mahdotonta havaita, ellei niitä olla tehty julkisiksi. Koska ongelma on nykyisellään ratkaisematon, on tärkeää tuoda käyttäjille selväksi, että tällaista väärinkäyttöä voi tapahtua, mikä voi kannustaa varovaisuuteen sisältöä julkaistessa. Tämän lisäksi on tärkeää painottaa webin louhijoiden keskuudessa sitä, että louhinta ei saisi loukata yhdenkään yksilön oikeutta yksityisyyteen. [21]

5.2 Web-käytön louhinnan yksityisyys

Palvelun tarjoajan näkökulmasta monet web-käytön louhinnan osa-alueista ovat varsin potentiaalisia tapoja tuoda parempaa kokemusta käyttäjälleen tai kehittää pidemmälle palvelua. Moni varsin hyödyllinen analyysi voidaan tehdä havainnoimalla käyttäjien erilaisia tapoja ja käyttötietoja ja palvelun tarjoaja voi halutessaan kerätä kaikki louhintaan

tarvittava data itse. Kuitenkin yksilön näkökulmasta on huomioitavaa, että tämä voi loukata suuresti yksilön yksityisyyttä, koska käyttäjästä voidaan kerätä dataa ilman, että hän on edes tietoinen siitä. Palvelimet usein tallentavat esimerkiksi vain lokitiedostoja, minkä takia Tämän lisäksi, vaikka käyttäjä tietäisi datan keräämisestä, ei hän voi olla varma siitä, että miten se tulee käytetyksi. Käyttäjien on mahdotonta tietää, että miten palvelu todellisuudessa käyttää keräämiään tietoja, koska mitään tapaa ei ole seurata annettuja dataa tai tietää palvelun käyttäytymisestä. [22]

Tämän lisäksi web palvelun käyttäjä tunnistautuu internetin tarjoajalle IP-osoitteellaan. Tarjoajan ulkopuolelle IP-osoite ei leviä kuin muodossa, jota ei voida kohdentaa yhteen tiettyyn yksittäiseen käyttäjään, vaan se voidaan kohdentaa palvelun tarjoajan alueeseen. Kuitenkin tarjoajalla on mahdollisuus seurata yksittäisten IP-osoitteen käyttäytymistä webissä, minkä pohjalta voi internetin tarjoaja kohdentaa webissä käyttäytymistä joko yksittäiseen talouteen tai henkilöön. Koska käyttäjällä ei ole mitään tapaa piilottaa IP-osoitetta palvelun tarjoajaltaan, ei käyttäjällä ole mahdollista käyttää internetiä ilman, että halutessaan internetin tarjoaja voisi seurata hänen käyttäytymistään. Joidenkin maiden lakien mukaan on kiellettyä palvelujen tarjoajan seurata yksittäisten osoitteiden toimintaa tai tallentaa niitä myöhempää käyttöä varten, ellei käyttäjän ja tarjoajan välinen sopimus sitä vaadi. Kuitenkin kaikissa maissa ei ole lakeja tai säädöksiä tähän liittyen, minkä takia on epäselvää, että kuinka moni tarjoaja väärinkäyttää asemaansa käyttäjiin nähden, ja rikkoo näin käyttäjän yksityisyyttä. [21]

Tähän eettiseen ongelmaan ei ole yksiselitteistä vastausta. Web-käytön louhinnasta on selkeästi monia hyödyllisiä ja täysin eettisesti päteviä käyttökohteita, minkä takia sen toteuttaminen on perusteltua. Yksilön suojaksi on olemassa monia eri työkaluja. Kolmannen osapuolen tuottamia työkaluja suojellakseen käyttäjiä on tullut markkinoille webin yleistyessä arkikäytössä. Monet niistä antavat enemmän vaihtoehtoja käyttäjälleen pysyttäytyä anonyyminä nettisivuilla, minkä myötä he voivat suojella yksityisyyttään. Työkalusta ja sen toteuttajasta riippuen kuitenkin ne eivät voi antaa täydellistä suojaa esimerkiksi internetin tarjoajalta tai voivat jo itsessään olla yksityisyyttä rikkovia tahoja. Työkalujen lisäksi on tärkeää informoida yksilöä hänen datastaan ja siitä, että miten hän voi halutessaan suojella yksityisyyttään. Sen lisäksi on tärkeää nostaa keskustelua siitä, että miten yksityisyyttä tulisi käsitellä internetissä, ja mihin käyttäjällä on oikeus. Yksikäsitteiset lait antaisivat selkeän vastauksen siihen, että miten internetin tarjoaja saa käsitellä asiakkaittensa dataa. Tämän lisäksi ne antaisivat selkeän kuvan siihen, että mitä palvelujen tarjoajat voivat taltioida käyttäjästä. Niin kauan, kun tällaisia lakeja ja säädöksiä ei ole määrittelemässä palvelun tuottajan toimintaa, on palvelun tuottajalla vapaus toteuttaa toimintaansa moraalinsa pohjalta. [21]

6. JOHTOPÄÄTÖKSET

Webin louhinta on tiedonlouhintaa, joka toteutetaan webistä saatavalla datalla. Se mahdollistaa webin suuren datamäärän hyödyntämisen. Erilaiset tiedonkeruumenetelmät yhdessä louhinnan eri menetelmät voivat tuottaa merkittävää analyysiä halutusta aiheesta tai kerätä tehokkaasti tietoa. Webin louhinnan avulla voidaan tutkia siis lähes mitä vain webin eri osa-alueita tarjoten lukemattomien eri sovelluksien tai analyysien toteuttamisen, jotka pohjautuvat webistä kerättyyn dataan.

Webin louhinta koostuu kolmesta eri osa-alueesta, joita ovat web-käytön louhinta, web-sisällön louhinta ja web-rakenteen louhinta. Nämä kolme osa-alueita eroavat sen perusteella toisistaan, että mitä osaa webistä ne tarkastelevat. Käyttö tarkastelee webin käyttäjiä ja niihin kohdennettua dataa, sisältö tarkastelee nettisivujen tarjoamaa sisältöä ja rakenne tarkastelee nettisivujen keskinäisiä suhteita. Kaikilla näillä osa-alueilla on eri käyttötarkoituksia ja erilaisia haasteita toteutuksessa. Kuitenkin usein niitä käytetään yhdessä saavuttaakseen monipuolisempaa analyysiä halutusta aiheesta.

Tunnetuimpia käyttökohteita webin louhinnalle on hakukone ja käyttäjälle kohdennettu sisältö. Hakukone voi hyödyntää kaikkia eri louhinnan osa-alueita. Käyttäjätietojen perusteella se arvioi eri hakutulosten korrelaation hakusanoihin, rakenteen louhimisen perusteella pystyy tarjoamaan analyysiä hakutulosten kontekstista ja omistussuhteista, ja sisällön louhinnan perusteella pystyy hakukone tuottamaan lyhyen koosteen nettisivun sisällöstä. Vastaavasti käyttäjälle kohdennettaessa sisältöä keskitytään enemmän käyttäjän tietoihin. Siinä tutkitaan sekä käyttäjän antamia tietoja että käyttäjän luomia tietoja hahmottaakseen mahdollisimman tarkasti sen, että millainen käyttäjä on kyseessä ja millaisesta tiedosta hän olisi mahdollisesti kiinnostunut.

Webin louhinta on itsessään eettisesti pitävää, jos se toteutetaan datalla, jota on lupa käyttää. Mutta on huomioitavaa, että koska käyttäjällä ei ole mitään tapaa hallinnoida hänestä lähtevää dataa, on hänen mahdotonta ylläpitää yksityisyyttään nykyaikaisista työkaluista huolimatta. Tämän takia käyttäjän dataa voi ajautua muiden käyttäjien tai palvelun tarjoajien toimesta ilman hänen tietoaan, mikä johtaa siihen, että data voi päätyä ilman hänen suostumustaan louhintaan. Tällöin louhinnan eettisyys on sen toteuttajan vastuulla. On huomioitavaa, että vaikka data olisikin täysin anonyymiä, on silti mahdollista, että käyttäjän yksityisyys tulee loukatuksi. Tämä johtuu siitä, että vaikka louhinnan tarkoituksena ei olisi kohdentaa tuotettua informaatiota yksittäiseen käyttäjään, voidaan datasta silti löytää yksilöiviä pisteitä.

LÄHTEET

- [1] Kosala, R. and Blockeel, H., 2000. Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1), pp.1–15.
- [2] Sharma, K., Shrivastava, G. and Kumar, V., 2011, April. Web mining: Today and tomorrow. In *2011 3rd International Conference on Electronics Computer Technology* (Vol. 1, pp. 399-403). IEEE.
- [3] Roiger, R.J., 2017. *Data mining: a tutorial-based primer*. Chapman and Hall/CRC.
- [4] Berson, A. and Smith, S.J., 1997. *Data warehousing, data mining, and OLAP*. McGraw-Hill, Inc.
- [5] Bramer, M., 2007. *Principles of data mining* (Vol. 180). London: Springer.
- [6] Srivastava, A., Desikan, P., and Kumar, V., 2005. Data Mining: Concepts, Models, Methods, and Algorithms. *Journal of Computing and Information Science in Engineering*, 5(4), pp.394–395.
- [7] Etzioni, O., 1996. The World-Wide Web: quagmire or gold mine? *Communications of the ACM*, 39(11), pp.65-68.
- [8] Cooley, R., Mobasher, B. and Srivastava, J., 1997, November. Web Mining: Information and Pattern Discovery on the World Wide Web. In *ictai* (Vol. 97, pp. 558–567).
- [9] Chen, H. and Chau, M., 2004. Web mining: Machine learning for web applications. *Annual review of information science and technology*, 38(1), pp.289–329.
- [10] Kumar, A. and Singh, R.K., 2016. Web mining overview, techniques, tools and applications: A survey. *International Research Journal of Engineering and Technology (IRJET)*, 3(12), pp.1543–1547.
- [11] Johnson, F. and Gupta, S.K., 2012. Web content mining techniques: a survey. *International Journal of Computer Applications*, 47(11).
- [12] Liu, B. and Chen-Chuan-Chang, K., 2004. Special issue on web content mining. *Acm Sigkdd explorations newsletter*, 6(2), pp.1–4.
- [13] Da Costa, M.G. and Gong, Z., 2005, June. Web structure mining: an introduction. In *2005 IEEE International Conference on Information Acquisition* (pp. 6–pp). IEEE.
- [14] Kumar, R. and Singh, A.K., 2010. Web structure mining: Exploring hyperlinks and algorithms for information retrieval. *American Journal of applied sciences*, 7(6), pp.840–845.
- [15] Fürnkranz, J., 2002. *Web Structure Mining. Exploiting the Graph Structure of the World-Wide Web*, Österreichische Gesellschaft für Artificial Intelligence (ÖGAI), pp.17–26.

- [16] Parvatikar, S. and Joshi, B., 2014. Analysis of user behavior through web usage mining. *Int. J. Comput. Appl*, pp.09750-8887. Parvatikar, S. and Joshi, B., 2014. Analysis of user behavior through web usage mining. *Int. J. Comput. Appl*, pp.09750–8887.
- [17] Cooley, R.W. and Srivastava, J., 2000. *Web usage mining: discovery and application of interesting patterns from web data*. Minneapolis, MN: University of Minnesota.
- [18] Eirinaki, M. and Vazirgiannis, M., 2003. Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)*, 3(1), pp.1–27.
- [19] Google Inc, www.google.com.
- [20] Pierrakos, D., Paliouras, G., Papatheodorou, C. and Spyropoulos, C.D., 2003. Web usage mining as a tool for personalization: A survey. *User modeling and user-adapted interaction*, 13(4), pp.311–372.
- [21] Van Wel, L. and Royakkers, L., 2004. Ethical issues in web data mining. *Ethics and Information Technology*, 6(2), pp.129–140.
- [22] Wang, Y., 2000. Web mining and knowledge discovery of usage patterns. *Cs 748T Project*, pp.1–25.