Tampereen yliopisto

Tommi Piili

MUSTAHATTUHAKUKONEOPTIMOINTI

Informaatioteknologian ja viestinnän tiedekunta
Pro gradu -tutkielma
Huhtikuu 2023

TIIVISTELMÄ

Tommi Piili: Mustahattuhakukoneoptimointi
[Pro gradu -tutkielma]
Tampereen yliopisto
Tietojenkäsittelytieteiden tutkinto-ohjelma
Huhtikuu 2023

Hakukoneoptimoinnin tarkoituksena on lisätä verkkosivun näkyvyyttä hakukoneiden tulossivuilla. Mustahattuhakukoneoptimointi on hakukoneyhtiöiden laatimien ohjesääntöjen vastaisten hakukoneoptimointimenetelmien hyödyntämistä. Tämän tutkielman tavoitteena on selvittää kirjallisuuskatsauksen pohjalta, mitä mustahattuhakukoneoptimoinnin menetelmiä ja vastatoimia tieteellisessä kirjallisuudessa on tutkittu. Tutkimusstrategiaksi valittiin kirjallisuuskatsaus, jotta tutkittava informaatio perustuisi tieteelliseen aineistoon. Kirjallisuuskatsauksen systemaattinen haku suoritettiin tietyin hakuehdoin neljään tietojenkäsittelytieteiden alan tietokantaan, minkä lisäksi suoritettiin täydentäviä lisähakuja kahteen muuhun tietokantaan. Tutkielman tavoitteena oli sisällyttää kirjallisuuskatsaukseen vähintään neljäkymmentä lähdettä sopivan laajuuden saavuttamiseksi. Aineisto valikoitui relevanssiin perustuen, joka arvioitiin tutkimalla artikkeleiden tiivistelmä ja tekemällä yleiskatsaus artikkeleiden sisältöön. Mustahattumenetelmät jaetaan sivun sisäisiin menetelmiin, linkkiperustaisiin menetelmiin ja muihin menetelmiin sekä tehostusmenetelmiin ja piilotusmenetelmiin. Lisäksi mustahattumenetelmiä voidaan hyödyntää verkkosivun hakutulossijoituksen tahalliseen alentamiseen. Liiallinen avainsanojen käyttö ja cloaking-menetelmä esiintyvät usein kirjallisuudessa. Mustahattuhakukoneoptimointi kuluttaa hakukoneiden resursseja, aiheuttaa verkkohakutuloksien laadun heikkenemistä ja voi edistää haitallisen verkkosisällön leviämistä. Hakukoneyhtiöt voivat antaa mustahattumenetelmiä hyödyntävälle verkkosivulle varoituksen, heikentää verkkosivun hakutulossijoitusta, tai poistaa verkkosivun hakuindeksistä. Mustahattuhakukoneoptimointia hyödyntävän verkkosivun toiminta voidaan pyrkiä lopettamaan myös oikeusteitse. Automaattiset menetelmät tehostavat mustahattuhakukoneoptimointia hyödyntävien verkkosivujen havainnointia. Mustahattumenetelmien kehittyessä myös vastatoimien on kehityttävä, jotta mustahattuhakukoneoptimoinnin vaikutuksia voidaan vähentää.

Avainsanat: Hakukoneoptimointi, Search engine optimization, SEO, Mustahattuhakukoneoptimointi, Spämmi, Spam

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

1 Johdanto

Hakukoneoptimoinnin (Search engine optimization, SEO) tarkoituksena on lisätä verkkosivun näkyvyyttä hakukoneiden tulossivuilla (Duk et al., 2013). *Mustahattuhakukoneoptimointi* on hakukoneyhtiöiden laatimien ohjesääntöjen vastaisten hakukoneoptimointimenetelmien hyödyntämistä (Fox, 2008; Jha & Saraswat, 2018; Li, 2014). Tämän tutkielman tavoitteena on selvittää kirjallisuuskatsauksen pohjalta, mitä mustahattuhakukoneoptimoinnin menetelmiä ja vastatoimia tieteellisessä kirjallisuudessa on tutkittu. Tässä tutkielmassa käytetään mustahattuhakukoneoptimoinnin menetelmistä termiä *mustahattumenetelmä.* Mustahattumenetelmiä hyödyntävään verkkosivuun viitataan tässä tutkielmassa termillä *mustahattusivu.* Aineistona on käytetty tieteellisiä julkaisuja. Menetelmiä ja niiden toimintalogiikkaa on pyritty tutkimaan yleistajuisesti, ilman syventymistä esimerkiksi algoritmien laskukaavoihin. Tavoitteena on saavuttaa rajatun aineiston puitteissa yleiskuva mustahattumenetelmien kirjosta, niiden toiminnasta ja vaikutuksista sekä mustahattuhakukoneoptimoinnin vastatoimista.

Tutkielman luvussa 2 kuvaillaan yleisesti tiedonhakua, tiedonhakuun liittyviä käsitteitä sekä hakukoneita. Luvussa 3 kuvaillaan yleisesti hakukoneoptimointia, joka jaetaan sivun sisäiseen ja ulkoiseen hakukoneoptimointiin sekä valkohattuiseen, mustahattuiseen ja harmaahattuiseen hakukoneoptimointiin. Luvussa 4 kuvaillaan yleisesti mustahattuhakukoneoptimointia, sen vastatoimia sekä negatiivista hakukoneoptimointia. Luvussa 5 kuvaillaan tutkielmassa käytetty tutkimusmenetelmä. Tutkimusmenetelmänä on kirjallisuuskatsaus. Luvussa esitellään kirjallisuuskatsauksessa käytetyt tietokannat sekä aineiston valikoitumisen perusteet. Luvussa 6 kuvaillaan sivun sisäisiä mustahattumenetelmiä ja niiden vastatoimia. Luvussa 7 kuvaillaan linkkiperustaisia mustahattumenetelmiä ja niiden vastatoimia. Luvussa 8 kuvaillaan cloaking- ja uudelleenohjausmenetelmiä sekä niiden havainnointia. Luvussa 9 kuvaillaan mustahattuhakukoneoptimoinnin vastatoimia ja niihin liittyviä haasteita. Luvussa 10 esitetään tutkielman tulokset. Luvussa 11 on keskusteluosuus, jossa reflektoidaan tutkimusmenetelmää ja kirjallisuutta sekä pohditaan mahdollisia jatkotutkimusmahdollisuuksia. Luvussa 12 on tutkielman yhteenveto.

2 Tiedonhaun menetelmät

Tässä luvussa kuvaillaan yleisesti tiedonhakua ja tiedonhakuun liittyviä tutkimusongelmia ja menetelmiä. Luvussa kuvaillaan myös hakukoneita. Alaluvussa 2.1 käsitellään tiedonhaun määritelmää, tiedonhaun osa-alueita, käsitteitä sekä menetelmiä. Alaluvussa 2.2 käsitellään vektorimallia ja TF-IDF-arvoa. Alaluvussa 2.3 käsitellään hakukoneiden määritelmää, hakukoneiden tehtäviä ja ongelmia.

2.1 Tiedonhaku

Tiedonhaussa tutkitaan informaation *esitystä* (representation), *organisointia* (organization), *tallennusta* (storage) ja *etsimistä* (accessing of information items) (Salton & McGill, 1987). Tiedonhakuun viitataan usein lyhenteellä *IR*, joka tulee tiedonhaun englanninkielen termistä *information retrieval* (Croft et al., 2009). Tiedonhaun yksi yleisimmistä sovellusalueista on *verkkohaku* (web search). *Vertikaalinen haku* (vertical search) on yksi verkkohaun muoto, jossa haun *verkkotunnus* (domain) on rajoitettu tiettyyn aiheeseen. Muita tiedonhaun muotoja ovat *yrityshaku* (enterprice search), *työpöytähaku* (desktop search) ja *vertaisverkkohaku* (peer-to-peer search, tai P2P). Yrityshaussa tiedonhaku rajoittuu yrityksen *sisäverkkoon* (intranet) ja työpöytähaussa yhden tietokoneen sisältöön. Vertaisverkkohaussa tiedonhaku kohdistuu verkkoon, jota ei ole kontrolloitu keskitetysti. (Croft et al., 2009)

Dokumentit ovat olleet tiedonhaun keskeisenä tutkimuskohteena 1950-luvulta lähtien. Erilaisia dokumentteja ovat esimerkiksi verkkosivut, sähköpostit ja kirjat. Dokumenteille on tyypillistä rakenne, johon dokumentin tyypistä riippuen voi sisältyä esimerkiksi otsikko, tekijä, päiväys ja tiivistelmä. Dokumentti koostuu yleisimmin tekstistä, joka voi rakenteettomuudestaan johtuen aiheuttaa haasteita saatettaessa sitä kuvailtavaan ja vertailtavaan muotoon. Dokumentti voi sisältää myös esimerkiksi kuvaa ja ääntä, joiden yhteyteen liitetään usein tekstikuvaus niiden sisällöstä. (Croft et al., 2009)

Tiedonhakumalli (information retrieval model) on sen prosessin muodollinen esitys, joka etsii vastaavuuden tiedonhaun yhteydessä esitetyn *kyselyn* (query) ja dokumentin välillä. Tutkijat hyödyntävät tiedonhakumalleja prosessien esittämiseen ja testaamiseen. Tiedonhaun eri tehtäviä ovat esimerkiksi *ad hoc -haku*, *suodatus* (filtering), *luokittelu* (classification) sekä kysymyksiin vastaaminen (question-answering). Tiedonhaun tärkeimpiä tutkimusongelmia ovat *relevanssi* ja *evaluointi,* eli miten hakutuloksia voidaan järjestellä tehokkaasti ja miten hakutuloksien relevanssia voidaan testata ja mitata. Lisäksi tiedonhaun tutkimusongelmiin kuuluu *vuorovaikutus käyttäjän kanssa* (user interaction) ja tiedonhaun tarpeiden tutkiminen. (Croft et al., 2009)

Relevanssi on tiedonhaun keskeinen käsite. Relevantti dokumentti voidaan yksinkertaisesti kuvailla niin, että se sisältää sen informaation, jota tiedonhakija etsi syöttäessään kyselyn. Ei ole kuitenkaan ongelmatonta tuottaa hakutulos, joka on hakijalle relevantti. Esimerkiksi kyselyssä esitettyä tekstiä täysin vastaavan tekstin etsiminen dokumentista voi tuottaa epärelevantin tuloksen. Croft ja muut (2009) viittaavat tähän ongelmaan termillä *vocabulary mismatch.* Vaikka hakutuloksena saadun dokumentin aihe vastaakin kyselyä, se ei välttämättä vastaa tiedonhakijan henkilökohtaisia tarpeita. Jos esimerkiksi kyselyn ”severe weather events” tuottama hakutulos on uutinen, joka käsittelee Kansasin osavaltiossa sattunutta tornadoa, tulos on aiheeltaan relevantti. Uutinen voi kuitenkin olla

vanha, eikä tiedonhakija välttämättä asu kyseisellä alueella, joten hakutulos ei ole hakijalle relevantti. Croft ja muut (2009) viittaavat tähän ongelmaan termeillä *aiherelevanssi ja käyttäjärelevanssi* (topical relevance and user relevance). (Croft et al., 2009)

Evaluointi on hakutuloksen relevanssin arviointia. Evaluoinnissa hyödynnetään mittareita, kuten *tarkkuus ja saanti* (precision and recall). Tarkkuus on relevanttien dokumenttien osuus kaikista hakutuloksina esitettävistä dokumenteista. Saanti on haettujen relevanttien dokumenttien osuus kaikista relevanteista dokumenteista. Saanti-mittarin käyttöön liittyy olettamus, että kaikki mahdolliset relevantit dokumentit ovat tiedossa. Näin ollen saanti toimii suljetussa testikokoelmassa. (Croft et al., 2009)

2.2 Vektorimalli ja TF-IDF

Vektorimalli (vector space model) esittää dokumentit ja kyselyt vektoriavaruudessa, joka havainnollistaa esimerkiksi dokumentin sisältämien *termien painoarvon* (term weighting) (Croft et al., 2009). Vektorimalli on klassinen tiedonhakumalli, joka mahdollistaa dokumenttien samankaltaisuuden vertailun. Sen avulla pyritään löytämään tiedonhakijan kyselyä vastaava relevantein dokumentti. (Poulimenou et al., 2016) Termien painoarvon selvittämisen lisäksi vektorimallia voidaan hyödyntää tekstin semanttiseen prosessointiin. Koska tietokoneet eivät ymmärrä kielen merkityksiä, ihmisen on haastavaa ohjeistaa tietokoneita, mikä aiheuttaa haasteita tekstin automaattiseen prosessointiin ja analysointiin. (Turney & Pantel, 2010)

TF-IDF (Term Frequency – Inverse Document Frequency) ilmaisee termien painoarvon (Croft et al., 2009). TF-IDF-arvoa pidetään yhtenä yleisimmin käytetyistä termien painoarvon mittareista tiedonhaussa (Poulimenou et al., 2016). TF-arvo ilmaisee tietyn termin esiintyvyyden dokumentissa (Croft et al., 2009). Esiintymistiheys ilmaisee, kuinka informatiivinen kyseinen termi on dokumentin kannalta (Ghosh & Desarkar, 2018). IDF-arvo ilmaisee termin esiintyvyyden koko kokoelman dokumenttien joukossa. IDF-arvon käänteisyys (inverse) viittaa siihen, että se antaa painoarvoa niille termeille, jotka esiintyvät harvoissa dokumenteissa. (Croft et al., 2009) IDF-arvo ilmaisee, kuinka informatiivinen kyseinen termi on dokumenttikokoelmassa. Mitä useammassa kokoelman dokumentissa kyseinen termi esiintyy, sen vähemmän informatiivinen se on. TF-IDF-esityksessä dokumentti esitetään vektorina, jonka kentät (fields) vastaavat sanaston termejä. Kenttiin merkitty arvo vastaa kyseisen termin TF-IDF-arvoa. (Ghosh & Desarkar, 2018)

Dokumenttikokoelman dokumenttien luokittelun (eli aiheen määrittelyn) välineenä käytetään dokumenteissa esiintyviä sanoja. Luokittelussa on syytä kiinnittää huomio erityisesti sellaisiin sanoihin, jotka ovat kuvaavia dokumentin aiheelle. Dokumentissa eniten esiintyvät sanat eivät usein ole luokittelun kannalta relevantteja, sillä ne sisältävät *hukkasanoja* (stop words), kuten “the” ja “and”. Toisaalta sanat, kuten “albeit” ja “notwithstanding” ovat harvinaisia, mutta eivät myöskään ole dokumentin aiheen kannalta kuvaa-

via. Näin ollen hyödylliset harvinaiset sanat esiintyvät vain harvoissa kokoelman dokumenteissa. Dokumenttikokoelman kannalta harvinaiset sanat ovat dokumentin luokittelun kannalta relevantteja, kun ne esiintyvät tietyssä dokumentissa useasti. TF-IDF-arvo on siis korkea, kun tietty termi esiintyy vain harvassa kokoelman dokumentissa, mutta siinä dokumentissa missä se esiintyy, se esiintyy useasti. TF-IDF-kaavan IDF-komponentti vähentää niiden hukkasanojen painoarvoa, jotka esiintyvät monissa dokumenteissa. (Rajaraman & Ullman, 2011)

2.3 Hakukoneet

Hakukone on tiedonhakumalleja hyödyntävä laajoja tekstikokoelmia käsittelevä sovellus, jonka yksi muoto on *verkkohakukone* (web search engine) (Croft et al., 2009). Tässä tutkielmassa hakukoneella viitataan verkkohakukoneeseen. Hakukone on tietokanta, joka indeksoi verkkosivuja mahdollistaen niiden etsimisen (Malaga, 2008). Yleisimmät hakukoneisiin liittyvät tutkimusongelmat ovat osin samat kuin tiedonhaussa yleisesti. Tutkimusongelmia ovat esimerkiksi evaluointi, tehokkaat *järjestelyalgoritmit* (ranking algorithms) sekä vuorovaikutteisuus (Croft et al., 2009).

Hakukoneiden tarkoituksena on tuottaa laadukkaita hakutuloksia etsimällä relevanteimmat verkkosivut kyselyyn nähden ja esittämällä niistä tärkeimmät. Relevanssiin vaikuttaa esimerkiksi dokumentin ja kyselyn tekstien samankaltaisuus. Relevanssin lisäksi hakutuloksia määrittävä tekijä on verkkosivun *tärkeys* (importance), joka on kyselystä riippumaton ominaisuus. Mitä enemmän verkkosivuille on osoitettu ulkoisia linkkejä, sen tärkeämpi se on. (Gyongyi & Garcia-Molina, 2005)

Hakukoneiden suorittaman hakutuloksien järjestelyn pohjana käytettävät tiedonhakumallit hyödyntävät aiherelevanssia ja käyttäjärelevanssia. Käyttäjärelevanssilla on evaluoinnin kannalta suurempi merkitys. Näin ollen käyttäjä päättää viime kädessä sen, onko hakutulos relevantti. Koska tekstimuodossa suoritettavat haut ovat puutteellisia kuvaamaan hakijan tiedontarvetta, relevanttien tulosten lisäämiseksi käytetään täydentäviä tekniikoita, jotka hyödyntävät vuorovaikutusta käyttäjän kanssa ja laajentavat haun kontekstia. Tällaisia tekniikoita ovat esimerkiksi hakusuositukset, hakujen laajennukset ja palautteet tulosten relevanssista. Hakukoneet huomioivat evaluoinnissa myös esimerkiksi sen, kuinka monta kertaa tiettyä hakutulosta on painallettu hakutulossivulla (clickthrough). (Croft et al., 2009)

Ensimmäinen verkkohakukone oli Archie vuodelta 1990, joka haki FTP-tiedostoja (File Transfer Protocol). Archie-hakukonetta seurasi tekstipohjainen hakukone Veronica. Vuonna 1993 julkaistiin GNA (Global Network Navigator), jota seurasi vuonna 1994 Yahoo! ja WebCrawler sekä Altavista vuonna 1995. Nykyisin lukuisista saatavilla olevista hakukoneista tunnetuimmat ovat Google (julkaistu 1998), Yahoo! Ja Microsoft Bing. (Varsha et al., 2021)

Hakukoneen komponentit koostuvat kahdesta pääasiallisesta funktiosta, jotka ovat *indeksointiprosessi* (indexing process) ja *kyselyprosessi* (query process). Indeksointiprosessi kokoaa ne rakenteet, jotka mahdollistavat haun suorittamisen ja joita kyselyprosessi hyödyntää yhdessä hakijan tekemän kyselyn kanssa tuottaakseen dokumenttien relevanssin perusteella järjestetyn listan (ranked list of documents). Indeksointiprosessin kolme pääasiallista komponenttia ovat tekstin hankkiminen ja muuntaminen sekä indeksin luonti. Indeksin luontiin liittyy dokumentissa esiintyvien termien painoarvon mittaaminen, joka voidaan ilmaista TF-IDF-arvona. (Croft et al., 2009)

Hakukoneet voidaan luokitella perustuen niiden toimintaperiaatteeseen. *Ryömijöitä* hyödyntäviin (crawler based) hakukoneisiin lukeutuvat ainakin Google, Yahoo!, Baidu, Yandex ja Bing. Niiden toiminta perustuu viiteen perusosaan, jotka ovat ryömintä, indeksointi, järjestely, prosessointi ja tiedon noutaminen tietokannasta (retrieving results). Toinen toimintaperiaate perustuu *manuaaliseen* toimintaan (human powered directories tai open directory systems). Haku voidaan esimerkiksi suorittaa tutkimalla verkkosivuston ylläpitäjän sivulleen kirjoittamaa metakuvausta, johon perustuva arvio tehdään manuaalisesti. *Hybridihakukoneet* (hybrid search engines) ovat yhdistelmä kahta edellä kuvattua toimintaperiaatetta. Esimerkiksi Googlen hakukone käyttää ensisijaisesti ryömijää, mutta toisinaan myös manuaalista indeksointia. (Varsha et al., 2021) Tässä tutkielmassa käsitellään ensisijaisesti ryömijöitä hyödyntäviä hakukoneita.

Ryömijöitä hyödyntävän hakukoneen osat ovat *pääte* (search interface), verkko (web), *kyselymoottori* (query engine), *hakuindeksi* tai *hakutietokanta* (search index), *indeksoija* (indexer) ja ryömijä (Levene, 2010). Ryömijä, johon viitataan myös termeillä *bot* tai *spider*, on ohjelmisto, joka etsii ja lataa internetissä saatavilla olevia verkkosivuja. Ryömijöiden keräämä sisältö lisätään hakukoneen tietokantaan. (Varsha et al., 2021) Ryömijät seuraavat verkossa olevia linkkejä löytääkseen uusia verkkosivuja ja pitääkseen jo ennalta indeksoidut verkkosivut ajan tasalla (Malaga, 2008). Indeksoija muodostaa ryömijöiden keräämästä informaatiosta hakutietokannan (Levene, 2010). Tätä prosessia kutsutaan indeksoinniksi. Hakutietokantaa voidaan kutsua myös *indeksiksi.* (Varsha et al., 2021) Hakutietokanta sisältää kaiken ryömijöiden keräämän informaation, jota hakukone tarvitsee verkkosivujen vertailuun ja noutamiseen (match and retrieve). Tähän informaatioon sisältyy esimerkiksi jokainen verkkosivulla esiintyvä sana. Hakukone järjestää verkkosivujen sanat aakkosjärjestykseen ja kerää jokaisen sanan viitetietoihin kaikki verkkosivut, joissa sana esiintyy. (Levene, 2010) Indeksiin tallennetaan myös sivun osoittamat ja sivulle osoitetut linkit. Indeksi sisältää suuren määrän verkkosivuja ja se luo verkkosivuista esitettävän mallin (representation) käyttämällä sivun URL-osoitetta ja verkkosivuun liittyvää sanaa tai lausetta. (Varsha et al., 2021) Kyselymoottori mahdollistaa hakujen suorittamisen ja hakutuloksien esittämisen. Se etsii ja järjestää hakulauseketta vastaavat hakutulokset tärkeysjärjestykseen. Järjestelyyn käytetään sivun vastaavuutta

hakulausekkeeseen, verkkosivulle osoitettujen linkkien määrää ja laatua sekä sivun suosiota. (Levene, 2010) Hakukoneet prosessoivat säilyttämäänsä dataa poistamalla epärelevantit sivut ja hukkasanat sekä normalisoimalla sananmuotoja (stemming). (Varsha et al., 2021)

Hakukoneet voidaan luokitella myös niiden tarjoamien palvelujen perusteella. *Tekstihakukoneet* (full text search engine) prosessoivat käyttäjän kyselyn ja etsivät verkosta kyselyä vastaavan relevantin informaation. Tekstihakukoneet voidaan edelleen jakaa kahteen alakategoriaan. Osa tekstihakukoneista noutaa hakutulokset omasta verkkotietokannastaan, kun taas osa käyttää ulkoista tietokantaa. *Hakemistohakukoneet* (directory search engine) etsivät informaation hakemistoista, jotka sisältävät tietyin ehdoin kategorisoituja verkkosivuja. *Metahakukoneet* taas suorittavat käyttäjän esittämän kyselyn useampaan hakukoneeseen ja listaavat tulokset yhteen paikkaan. *Vertikaaliset hakukoneet* ovat verkkotunnuskohtaisia ja käyttävät tiedonhakuun vain vähän resursseja. (Varsha et al., 2021)

Hakukoneiden suunnitteluun (search engine design) liittyy useita tutkimusongelmia. Yksi ongelma on hakukoneiden suorituskyky ja se, miten tiedonhakua ja indeksointia voidaan tehostaa. Toinen ongelma liittyy uuden datan sisällyttämiseen sekä datan kattavuuteen ja tuoreuteen. Datan ja käyttäjien mukana skaalautuminen sekä sopeutuminen (adaptability) aiheuttavat haasteita. Hakukoneisiin liittyy myös yksityiskohtaisempia ongelmia, kuten *spämmi* (spam), jota käsitellään luvussa 4.1. (Croft et al., 2009)

3 Hakukoneoptimointi

Tässä luvussa kuvaillaan yleisesti hakukoneoptimointia ja sen eri muotoja. Luvussa kuvaillaan hakukoneiden käsittelemän verkon sisältöä ja sitä, miksi korkeat hakutulossijoitukset ovat perusteltuja. Lisäksi luvussa kuvaillaan hakutuloksien järjestämisen perusteita. Alaluvussa 3.1 pyritään määrittelemään hakukoneoptimointi sekä kuvaillaan yleisesti hakukoneoptimointia ja sen perusteita. Alaluvussa 3.2 kuvaillaan verkkosivun rakenteellisia osia ja sivun sisäistä hakukoneoptimointia. Alaluvussa 3.3 kuvaillaan sivun ulkoista hakukoneoptimointia ja linkkiperustaisia järjestelyalgoritmeja. Alaluvussa 3.4 kuvaillaan hakukoneyhtiöiden ohjesääntöjä ja valkohattuista hakukoneoptimointia sekä verrataan valkohattuista hakukoneoptimointia mustahattuiseen hakukoneoptimointiin. Alaluvussa 3.5 kuvaillaan harmaahattuista hakukoneoptimointia.

3.1 Hakukoneoptimointi yleisesti

Hakukoneoptimointi on yksi *hakukonemarkkinoinnin* (search engine marketing, SEM) muoto. Search engine marketing professional organization (SEMPO) mukaan hakukonemarkkinointi on yksi internetmarkkinoinnin muoto. Hakukonemarkkinoinnin ja hakukoneoptimoinnin tarkoituksena on lisätä verkkosivujen näkyvyyttä hakukoneiden tulossivuilla (search engine result page – SERP). (Lynn et al., 2015; Duk et al., 2013) Tässä

tutkielmassa hakukoneiden tulossivusta käytetään termiä hakutulossivu ja siinä esiintyvistä tuloksista käytetään termiä hakutulokset. Hakukoneoptimoinnin tavoitteena on, että optimoitava verkkosivu sijoittuu korkeammalle yhden tai useamman hakukoneen hakutulossivulla tiettyä kyselyä kohden. Hakukoneoptimointimenetelmät voivat parantaa sivun sisällön laatua ja näin ollen lisätä sen saamaa arvostusta hakukoneyhtiöiden ja tiedonhakijan näkökulmista. Hakukoneoptimointi käsittää useita optimointimenetelmiä ja hakukoneyhtiöt tarjoavat työkaluja, analytiikkaa ja ohjeistuksia hakukoneoptimoinnin toteuttamiseen. (Roslina & Shahirah, 2019)

Verkkosivujen ja etenkin Web 2.0 sovelluksien lukumäärän kasvu on aiheuttanut internetissä saatavilla olevan informaation määrän nopean kasvun. Suuresta informaation määrästä johtuen relevantin informaation löytämiseen kuluu enemmän aikaa ja resursseja. Näin ollen hakukoneista on tullut pääasiallinen tiedonhaun väline internetissä. (Li, 2014)

Verkkosivut ovat yleisesti tärkeä osa liiketoimintaa. Ne ovat yrityksen näkyvä osa, jonka avulla yrityksistä etsitään tietoa. Yrityksen verkkosivujen korkea sijoittuminen hakukoneiden hakutulossivuilla on tärkeä keino tulla nähdyksi ja sivujen kävijämäärän kasvu voi lisätä yrityksen tuloja. (Duk et al., 2013; Kumar et al., 2016; Li, 2014) Hakutuloksissa korkealle sijoittuminen voi nostaa verkkosivulle kohdistuvien ulkoisten sivujen osoittamien linkkien määrää. Tämä taas voi edelleen johtaa sivun sijoittumisen nousuun, ja tuoda sivustolle mainoksia. (Kumar et al., 2016)

Googlen hakukone prosessoi kymmeniä tuhansia hakuja sekunnissa ja miljardeja hakuja päivässä. Suuri osa internetin käyttäjistä päätyy vierailemalleen verkkosivulle juuri hakukoneiden välityksellä. Usein hakukoneiden käyttäjät tutkivat korkeintaan ensimmäiset hakutulossivut (Liu et al., 2020). Ensimmäisten hakutulossivujen joukosta käyttäjät tutkivat erityisesti ensimmäisen hakutulossivun (Liu et al., 2020; Duk et al., 2013). Todennäköisimmin tiedonhakija valitsee yhden korkeimmalle sijoittuneista hakutuloksista (Li, 2014). Vain harvat valitsevat hakutuloksen, joka ei ole kolmen ensimmäisen hakutulossivun joukossa (Malaga, 2008). Useat käyttäjät tutkivat vain ensimmäisen hakutulossivun ja muuttavat kyselyään ennemmin kuin jatkavat toiselle hakutulossivulle (Gudivada et al., 2015). Usein käyttäjät valitsevat hakutuloksen, joka on kolmen korkeimmalle sijoittuneen tuloksen joukossa (golden triangle) (Varsha et al., 2021).

Hakukoneet käsittelevät valtaosaa verkon sisällöstä, joten hakutuloksien tehokas käsittely on tärkeää relevantin informaation etsimisen kannalta. Relevantin ja tarkan informaation erottaminen onkin hakukoneiden ensisijainen tarkoitus ja hakukoneoptimointi edistää tavoitteen saavuttamista. (Varsha et al., 2021) Eri hakukoneet järjestävät hakutulokset osittain toisistaan poikkeavin perustein. Esimerkiksi Google, Bing ja Yahoo! -hakukoneet arvostavat kaikki sivulle osoitettuja linkkejä, mutta sivun eri rakenteellisten osien painoarvo voi vaihdella hakukoneiden välillä. (Lynn et al., 2015) *PageRank* on algoritmi, jota esimerkiksi Googlen hakukone hyödyntää hakutuloksien järjestämiseen.

PageRank -algoritmin yhtenä järjestelyperusteena on, että mitä enemmän verkkosivulle on osoitettu ulkoisia linkkejä, sen korkeammalle se sijoittuu hakutuloksissa. (Killoran, 2013)

Hakukoneoptimoinnin tuloksena sijoittuvat hakutulokset ovat *luonnollisia hakutuloksia* (organic results), jolloin niiden sijoittumisen perusteena on sivun relevanssi hakijan kyselyyn nähden. Hakutulokset voivat olla myös maksettuja, jotka esiintyvät yleensä hakutulossivun yläosassa tai oikeassa laidassa. (Varsha et al., 2021)

Sivun sisällön laadun parantaminen on yksi hakukoneoptimoinnin keino. Laatu voi olla kuitenkin vaikeasti määriteltävissä, mutta yleensä laatuun liitetään käsitteitä, kuten *käytettävyys*, *luotettavuus* ja *turvallisuus*. Hakukoneoptimointimenetelmiin liittyy pääasiassa viisi attribuuttia, jotka määrittävät sivun laatua. Nämä ovat *virheettömyys* (correctness), esitys (presentation), sisältö, *navigointi* ja vuorovaikutus. Yleisesti ottaen hakukone tutkii verkkosivun käytettävyyttä, sijoituksen korkeutta, sisällön rikkautta, saavutettavuutta (kuinka nopeasti sivu latautuu sekä sivun luettavuus ja ymmärrettävyys), kuvien ja videoiden tai muiden sisällön muotojen sisällyttämistä, tyhjentävyyttä ja asiaankuuluvuutta, päivityksiä (onko sivu päivitetty tarvittaessa tai kun se on ollut mahdollista), turvallisuutta sekä sivulle osoittavien linkkien laatua ja määrää. Nämä laadulliset attribuutit luovat pohjan hakukoneoptimoinnille, jotta sen metodit voidaan määritellä tarkemmin ja selkeämmin. (Varsha et al., 2021)

Dukin ja muiden (2013) mukaan Googlen hakukone hyödyntää yli kahtasataa signaalia hakutuloksien järjestämiseen. Hakukoneiden algoritmit muuttuvat jatkuvasti, mikä vaikeuttaa hakukoneen toiminnan saamista selville (Wynne, 2012). Sen lisäksi hakukoneyhtiöt pitävät hakukoneidensa tarkat toimintaperiaatteet salaisina (Levene, 2010).

Hakukoneoptimointi on vaiheittaista ja kattaa eri osa-alueita sivun asettelusta yksittäisiin elementteihin. Hakukoneoptimointi on jatkuva prosessi, joka vaatii sivustovastaavalta optimointimenetelmien ja sivun sisällön päivittämistä. Hakukoneoptimointi ei liity vain verkkosivun attribuutteihin, vaan näkyvyyttä voi edistää myös esimerkiksi mainostamalla sosiaalisessa mediassa. Digitalisaation johdosta edistyneiden hakukoneoptimointimenetelmien omaksumisen merkitys kasvaa markkinoinnin ja liiketoiminnan kannalta. Hakukoneoptimointi on tärkeää myös hakukoneen käyttäjän luottamuksen kannalta ja mahdollistaa sivun kävijäliikenteen kestävän kasvun. (Varsha et al., 2021)

Hakukoneoptimointimenetelmät voidaan jakaa sivun sisäiseen ja ulkoiseen optimointiin. Sivun sisäisessä hakukoneoptimoinnissa pyritään optimoimaan esimerkiksi sivun sisältötekstiä ja metadataa, kun taas sivun ulkoinen hakukoneoptimointi liittyy sivulle osoittaviin linkkeihin. (Wynne, 2012)

3.2 Sivun sisäinen hakukoneoptimointi

Sivun sisäiset hakukoneoptimointimenetelmät kohdistuvat sivun sisäisiin osiin. Verkkosivun keskeisiä osia ovat *HTML, CSS* ja *JavaScript,* jotka ovat verkkosivujen kehityksessä hyödynnettäviä merkintä- ja ohjelmointikieliä. HTML määrittelee sivun sisällön, CSS tyylin ja JavaScript toiminnallisuuden. (w3schools.com[1]) Myöhemmin tässä luvussa kuvaillaan, miten sivun sisäiset hakukoneoptimointimenetelmät kohdistuvat sivun HTML -osiin. Sivun CSS- ja JavaScript -osia voidaan hyödyntää mustahattuhakukoneoptimoinnissa, mitä kuvaillaan esimerkiksi luvuissa 6.3 ja 8.3.

HTML (Hypertext markup language) on verkkosivujen merkintäkieli, joka kuvailee verkkosivun rakennetta (w3schools.com[2]). Se koostuu elementeistä, jotka selain muuttaa sivun näkyväksi osaksi. Elementeistä käy ilmi esimerkiksi sivun otsikko, kappaleet ja linkit. Elementit merkitään *tunnisteilla* (tag), joihin kuuluu *aloitustunniste* (esimerkiksi <html>) ja lopetustunniste (esimerkiksi </html>). Elementtiin kuuluu kaikki aloitus- ja lopetustunnisteiden välissä oleva sisältö. Seuraa esimerkki HTML-tiedoston sisällöstä selityksineen (HTML-dokumentissa kommentit kirjoitetaan merkkien <!-- ja --> väliin):

```
<!DOCTYPE html> <!--määrittelee kyseessä olevan HTML5-dokumentti-->
<html> <!-- HTML-sivun juurielementti -->
<head> <!-- sisältää metatietoa HTML-sivusta -->
<title>Page Title</title> <!-- määrittelee HTML-sivun otsikon, joka
on näkyvillä selaimen otsikkorivillä tai sivun välilehdellä-->
</head>
<body> <!-- määrittelee dokumentin rungon, joka sisältää kaiken se-
laimen käyttäjälle näkyvän sisällön-->
<h1>My First Heading</h1> <!--määrittelee sisältöön kuuluvan otsikon
-->
<p>My first paragraph.</p> <!-- määrittelee tekstikappaleen -->
</body> </html>
```

(w3schools.com[3])

CSS (Cascading style sheets) on kieli, joka määrittelee, miltä HTML-elementit näyttävät. CSS-merkinnät voidaan sisällyttää CSS-dokumenttiin. Esimerkiksi HTML <body> -elementin tyyli voidaan määritellä seuraavasti:

```
body {background-color: lightblue;}
```
(w3schools.com[4])

[1] *https://www.w3schools.com/default.asp*, noudettu 16.3.2023
[2] *https://www.w3schools.com/html/default.asp*, noudettu 16.3.2023
[3] *https://www.w3schools.com/html/html_intro.asp*, noudettu 16.3.2023
[4] *https://www.w3schools.com/css/default.asp*, noudettu 16.3.2023

JavaScript on verkkoympäristössä käytettävä ohjelmointikieli, jonka avulla verkkosivulle voidaan sisällyttää dynaamista sisältöä (w3schools[5]). Seuraa esimerkki, joka avulla verkkosivulle voidaan lisätä kellonajan ja aikavyöhykkeen ilmaiseva painike:

```
<button type="button"
onclick="document.getElementById('demo').innerHTML = Date()">
Click me to display Date and Time.</button>
<p id="demo"></p>
```

(w3schools.com[6])

Sivun sisäiset hakukoneoptimointimenetelmät ovat suoraan sivustovastaavan hallinnassa ja ne ovat merkittävimpiä ja eniten käytettyjä hakukoneoptimointikeinoja (Varsha et al., 2021). Sivun sisäiselle hakukoneoptimoinnille on ominaista, että optimoitavan sivuston sivut, otsikot, tunnisteet, sisältö ja rakenne ovat selkeitä ja sisältävät relevantteja *avainsanoja* (Roslina & Shahirah, 2019). Sivun tekstikenttiä sisältävät osat ovat yleensä dokumentin *runko* (document body), *title-elementti*, *metatunnisteet* (meta tag), *HTML-header* ja URL-osoite. Myös sivulle osoittavan linkin *ankkuriteksti* voidaan katsoa kuuluvan sivun tekstikentäksi. (Gyongyi & Garcia-Molina, 2005) Muita sivun sisäisen hakukoneoptimoinnin kannalta tärkeitä osia ovat *title-tunnisteet* (title tag)*,* metakuvaukset*, meta-avainsanat* sekä *alt-tunnisteet.* (Roslina & Shahirah, 2019)

Sivun sisäisessä hakukoneoptimoinnissa tärkein menetelmä on *avainsanojen* käyttö (Duk et al., 2013). Sivun hakukoneoptimoinnista vastaavan tulisi suosia sellaisia avainsanoja ja -lauseita, jotka liittyvät sellaiseen ongelmaan tai tarpeeseen, joita sivun kohderyhmään kuuluva tiedonhakija todennäköisesti käyttää hakusanoinaan (Killoran, 2013). Avainsanojen tulisi aiheeltaan sivun sisällölle kuvaavia (Jha & Saraswat, 2018; Varsha et al., 2021; Patil et al., 2021), jotta hakukoneet pystyvät niiden perusteella sijoittamaan sivun tiettyyn kategoriaan (Varsha et al., 2021). Osuvien hakusanojen tai hakulausekkeiden keksiminen ei ole välttämättä helppoa, sillä kohderyhmään kuuluvien tiedonhakijoiden hakusanat eivät välttämättä vastaa asiantuntijoiden käyttämiä hakusanoja (Killoran, 2013).

Avainsanoja tulisi sijoitella sivun tärkeisiin osiin, kuten kuvien *alt-teksteihin*, title-elementteihin (Gandour & Regolini, 2011), URL-osoitteeseen, metatunnisteisiin, ankkuriteksteihin (Varsha et al., 2021) ja sisältötekstiin (Zhang & Dimitroff, 2005). Sivun URL-osoitteen tulisi olla lyhyt ja ymmärrettävä ja sen tulisi sisältää avainsanoja. Sivun metatunniste sisältää lyhyen kuvauksen sivun sisällöstä. Hakukoneiden ryömijät hyödyntävät

[5] *https://www.w3schools.com/js/default.asp*, noudettu 16.3.2023
[6] *https://www.w3schools.com/js/tryit.asp?filename=tryjs_myfirst*, noudettu 16.3.2023

erityisesti metatunnisteessa olevia avainsanoja indeksoinnin perusteena. Alt-teksti on lyhyt kuvaus sivun sisältämistä kuvista ja videoista. Hakukone indeksoi sivun kaikki objektit, joten lyhyt kuvaus kuvien tai videoiden sisällöstä avainsanoja hyödyntäen on suositeltavaa. Ankkuriteksti on toiseen dokumenttiin osoittavassa linkissä oleva teksti. Sivustovastaavien on suositeltavaa käyttää avainsanoja ankkuritekstissä nostaakseen sivun sijoitusta. (Varsha et al., 2021) Avainsanojen sisällyttäminen myös sivun metatietoihin on kannattavaa, sillä metatietoja hyödyntäen hakukoneet keräävät informaation sivun sisällöstä (Roslina & Shahirah, 2019).

Avainsanatiheys (keyword density) on avainsanojen käyttöä rajoittava tekijä (Varsha et al., 2021). Avainsanatiheys on avainsanojen prosentuaalinen osuus koko sivun sanamäärästä (Zuze & Weideman, 2013). Aihetta tutkivassa kirjallisuudessa on esiintynyt keskustelua avainsanatiheyden merkityksestä verkkosivun sijoittumisessa hakutuloksissa. Intuitiivisesti voisi päätellä, että mitä enemmän avainsanoja on, sitä korkeammalle sivu sijoittuu. Hakukone voi kuitenkin tulkita avainsanojen liiallisen käytön pyrkimykseksi vaikuttaa hakutuloksiin. Jos hakukone tulkitsee verkkosivun käyttävän avainsanoja väärin, se voi heikentää sivun sijoittumista tai jopa poistaa sen hakutuloksista. (Zuze & Weideman, 2013). Sopivasta avainsanatiheydestä ei ole kirjallisuudessa selkeää yksimielisyyttä. Suositellut avainsanatiheydet ovat vaihtelevat 2-8 prosentin välillä (Jha & Saraswat, 2018; Varsha et al., 2021). Sivustovastaavien tulisi välttää sivujen ylioptimointia, eikä luoda avainsanoja automaattisesti (Patil et al., 2021).

Sivun sisäinen hakukoneoptimointi kohdistuu myös verkkosivujen rakenteeseen. Hyvin optimoidun sivun tulisi olla rakenteellinen ja hyvin organisoitu. Sen tulisi sisältää hyvä navigointimahdollisuus sekä osuvat ja erottuvat otsikot. (Varsha et al., 2021) Sivun sisällön tulisi olla jaettuna segmentteihin käyttäen *heading-tunnisteita*. Heading-tunnisteiden tulisi sisältää avainsanoja ja kuvastaa sen alla olevaa sisältöä. (Jha & Saraswat, 2018).

Esimerkki metatunnisteen käytöstä HTML-tiedostossa:

```
<!DOCTYPE html>
<html>
<head>
<meta charset=”utf-8”>
<meta name=”Description” CONTENT=”The description of our meta tag”>
<meta name=”google-site-verification" content=”The Google site con-
sole”/>
<title>Enter the title of your website</title>
<meta name=”robots” content=”noindex,nofollow”>
```

(Roslina & Shahirah, 2019)

Sivun otsikon tulisi ilmaista sivun välittämän tuotteen tai palvelun tarkoituksen selkeästi ja sen ei tulisi olla liian pitkä. Otsikkoon tulisi sisällyttää osuvia avainsanoja ja sivun lukijoiden tulisi saada hyvä käsitys sivun sisällöstä vain lukemalla sen otsikon. Sivun URL-osoitteen tulisi olla selkeä ja sen pituuden tulisi olla lyhyt. Kohdistettu URL-osoite tehostaa ryömijöiden toimintaa. (Varsha et al., 2021)

Verkkosivun kokoon vaikuttavia tekijöitä ovat siihen liittyvät tiedostot, kuten HTML-tiedosto sekä visuaaliset ja upotetut objektit. Sivun koon ei tulisi olla liian suuri, jotta hakukoneet kykenevät tallentamaan verkkosivun tietokantaansa ja käyttäjän selain kykenee lataamaan sen nopeasti. Sivun latautumisaika, johon vaikuttaa esimerkiksi visuaalisten elementtien käyttö, vaikuttaa sen sijoittumiseen. (Varsha et al., 2021)

Sivun sisällön laadukkuus toimii myös hakukoneoptimointimenetelmänä. Laadukas sisältö on hyvin kirjoitettua sekä informatiivista, ja se vastaa hakijan tarpeita (Jha & Saraswat, 2018). Laadukkaan sisällön tulisi olla syvällistä, pitkällä aikavälillä relevanttia ja arvokasta sivun vierailijalle. Sivustovastaavan tulisi tarkistaa sivuston johdonmukaisuus, käytettävyys, relevanssi, luotettavuus ja auktoriteetti. Sivuston asiasisällön suhde oheissisältöön tulisi olla sopiva. Esimerkiksi mainoksia ei tulisi olla liikaa suhteessa sisältöön. Sivuston asiasisällön tulisi olla relevanttia käyttäjän verkkohakukyselyyn nähden. Sivustolta tulisi poistaa jäljennetty sisältö ja käyttää nofollow-komentoa. Mahdollisten ongelmallisten sisäisten sivujen indeksointia tulisi välttää, joiden sisältö on esimerkiksi jäljennettyä tai lähes jäljennettyä. (Patil et al., 2021)

3.3 Sivun ulkoinen hakukoneoptimointi

Sivun ulkoinen hakukoneoptimointi viittaa menetelmiin, jotka eivät suoraan liity optimoitavan sivun tai sivuston sisältöön. Ulkoisessa hakukoneoptimoinnissa on keskeistä *linkkien rakentaminen* (link building), joka viittaa muilta sivustoilta *kohdesivustolle* (optimoitavalle sivustolle) osoittavien linkkien kerryttämiseen. Kohdesivustolle osoittavista

linkeistä käytetään kirjallisuudessa termiä *back link* (Varsha et al., 2021; Patil et al., 2021), ja sivun osoittamista linkeistä termiä *forward link* (Patil et al., 2021)

Hakukoneyhtiöt arvostavat sivustoja, joille on osoitettu linkkejä, sillä se viittaa laadukkaaseen sisältöön ja luotettavuuteen. Ulkoisessa hakukoneoptimoinnissa yksi sijoitukseen vaikuttava tekijä on sivuston suosio (link popularity), joka viittaa siihen, kuinka monta linkkiä sivustolle on osoitettu muilta sivustoilta. (Varsha et al., 2021)

Mitä enemmän sivustolle on osoitettu ulkoisia linkkejä, sitä enemmän hakukoneyhtiöt arvostavat sivustoa. Linkkien laatu on kuitenkin tärkeämpi kuin määrä. (Killoran, 2013) Linkkejä osoittavien sivustojen tulisi sisältää laadukasta sisältöä ja hakulausekkeita vastaavia avainsanoja (link reputation) (Varsha et al., 2021). Hakukoneyhtiöt arvostavat linkkejä, jotka ovat osoitettu luotettavilta ja arvovaltaisilta sivustoilta (Wynne, 2012). Sijoittumista parantaa myös se, että linkit tulevat samaa aihetta käsitteleviltä sivustoilta. Myös linkkien ikä vaikuttaa niiden arvokkuuteen. (Killoran, 2013) Linkkejä osoittavien sivustojen tulisi kuitenkin olla sellaisia, ettei niitä ole poistettu hakukoneiden indeksistä tai hakukoneet eivät ole alentaneet niiden sijoittumista (Jha & Saraswat, 2018). Sivustolle on suositeltavaa laatia laadukas sisältö, jotta muut sivustot haluavat osoittaa linkkejä kyseiselle sivustolle (Killoran, 2013). Verkkosivustot, jotka sisältävät runsaasti ainutlaatuisia, muille sivustoille osoittavia linkkejä, sijoittuvat usein korkeammalle hakutuloksissa. Linkkien sijoittelu sosiaaliseen mediaan ja foorumeille (kuten Reddit) voi kasvattaa sivuston vierailijoiden määrää sekä sivuston mainetta. (Varsha et al., 2021)

Linkkiperustaisesti hakutuloksia järjestelevän *HITS-algoritmin* (hyperlink-induced topic search) suorittama laskelma relevanssista perustuu kahteen verkkosivulle laskettavaan arvoon: *hub* (hub score) ja *auktoriteetti* (authority score) (Manning et al., 2008; Patil et al., 2021; Gyongyi & Garcia-Molina, 2005; Wu & Davison, 2005). Mitä enemmän sivu osoittaa linkkejä sivustoille, joilla on suuri auktoriteettiarvo, sitä suurempi sen hub-arvo on. Mitä enemmän sivulle osoitetaan linkkejä sivuilta, joilla on suuri hub-arvo, sitä suurempi auktoriteettiarvo sillä on. HITS-algoritmi palauttaa hakutuloksena verkkohaun aiheeseen liittyvät hub- ja auktoriteettiarvojen perusteella järjestetyt sivut. (Gyongyi & Garcia-Molina, 2005; Wu & Davison, 2005)

Toinen linkkien perusteella hakutuloksia järjestelevä algoritmi on PageRank (Gyongyi & Garcia-Molina, 2005). Sivulle osoittavat linkit vaikuttavat olennaisesti sivun PageRank-arvoon (Patil et al., 2021). Mitä enemmän sivulle osoitetaan linkkejä ulkoisilta sivuilta, sitä suurempi sivun PageRank-arvo on. Myös linkin osoittavan sivun PageRank-arvo vaikuttaa linkin kohteena olevan sivun PageRank-arvoa nostattavaan vaikutukseen. (Gyongyi & Garcia-Molina, 2005) Jos sivulle osoitetaan runsaasti linkkejä arvokkailta sivuilta, kyseisen sivun linkillä osoittamien sivujen arvo nousee (Patil et al., 2021).

3.4 Valkohattuinen hakukoneoptimointi

Hakukoneyhtiöt ovat laatineet hakukoneoptimointiin liittyviä *ohjesääntöjä* (SEO Guidelines). Sääntöjä on runsaasti, mutta ohjesääntöjen pääprinsiippi voidaan tiivistää esimerkiksi seuraavasti: ”Tee uniikki, korkealaatuinen sivu, joka vastaa hakijan tarpeita. Pyri tekemään paras mahdollinen tulos relevanteille hauille, äläkä koeta saada tuloksia kyseenalaisin keinoin.” (Fox, 2008) Verkkosivu tulisi optimoida niin, että hakukoneiden ryömijöille esitettävä sisältö vastaa selaimen käyttäjälle esitettävää sisältöä (Malaga, 2008). Sivun sisällön tulisi olla alkuperäistä ja laadukasta (Varsha et al., 2021).

Kun hakukoneoptimointia toteutetaan ohjesääntöjen mukaisesti, hakukoneoptimointia voidaan pitää *valkohattuisena hakukoneoptimointina* (Jha & Saraswat, 2018; Li, 2014), tai *eettisenä hakukoneoptimointina* (Roslina & Shahirah, 2019). Ohjesääntöjen vastaista hakukoneoptimointia taas voidaan pitää mustahattuisena hakukoneoptimointina (Fox, 2008; Jha & Saraswat, 2018; Li, 2014).

Jotta verkkosivun sijoitus nousisi hakutuloksissa, valkohattuinen hakukoneoptimointi vaatii mustahattuista hakukoneoptimointia enemmän panostusta (Jha & Saraswat, 2018). Usein verkkosivujen ylläpitäjät kuitenkin harjoittavat hakukoneoptimointia hakukoneyhtiöiden laatimien sääntöjen mukaisesti (Li, 2014). Valkohattuinen hakukoneoptimointi on mustahattuista hakukoneoptimointia riskittömämpää (Varsha et al., 2021). Hakukoneyhtiöt eivät anna sanktioita havaitessaan ohjesääntöjen mukaista hakukoneoptimointia. Näin ollen riski siihen, että hakukoneyhtiöt hylkäävät verkkosivun hakutuloksistaan, tai heikentävät sen sijoittumista, on pienempi. (Jha & Saraswat, 2018; Roslina & Shahirah, 2019)

Hakutulossijoituksien kannalta valkohattuinen hakukoneoptimointi voi olla pitkällä aikavälillä mustahattuista hakukoneoptimointia kannattavampaa (Varsha et al., 2021). Valkohattumenetelmin saavutettu sijoituksen nousu tapahtuu hitaasti, mutta sijoitus voi kohota ajan saatossa korkeammalle kuin mustahattuisessa hakukoneoptimoinnissa. Lisäksi sivun kävijäliikenne säilyy pidempään. Sijoituksen noustua valkohattumenetelmin saavutettu tulos jatkaa kasvuaan, vaikka panostusta hakukoneoptimointiin laskettaisiinkin. Mustahattuinen hakukoneoptimointi mahdollistaa hakutulossijoituksien nopean kasvun, mutta se ei mahdollista saman tuloksen saavuttamista, kuin valkohattuinen hakukoneoptimointi parhaimmillaan. Mustahattumenetelmin saavutettu menestys on lyhytkestoista, ja voi hävitä lyhyessä ajassa kokonaan. (Jha & Saraswat, 2018)

Noudattamalla hakukoneyhtiöiden ohjesääntöjä ja verkkosivujen ylläpitäjät pyrkivät saavuttamaan hakukoneoptimoinnilla pitkäaikaista hyötyä. Myös hakukoneyhtiöt ja tiedonhakijat hyötyvät valkohattuisesta hakukoneoptimoinnista, sillä se edistää tiedonhakijan pääsyä laadukkaan sisällön ääreen hakukoneiden välityksellä. (Roslina & Shahirah, 2019) Taulukossa 1 kuvaillaan valkohattuisen ja mustahattuisen hakukoneoptimoinnin keskeisiä eroja.

Hakukoneoptimointimenetelmät	Valkohattuinen	Mustahattuinen
Määritelmä	Hakukoneyhtiöiden ohjesääntöjen mukainen hakukoneoptimointi. Menetelmiin on viitattu myös autenttisena tai eettisenä hakukoneoptimointina.	Menetelmillä pyritään kasvattamaan sivun sijoitusta hakutulossivulla hakukoneyhtiöiden ohjesääntöjen vastaisesti.
Orientaatio	Menetelmät ovat suunniteltu hyödyttämään ihmiskäyttäjää. Sisältö pyritään pitämään laadukkaana.	Menetelmien tavoitteena on harhauttaa hakukoneita käyttämällä epärelevantteja linkkejä ja sisältöä. Menetelmiä käytetään ensisijaisesti ryömijöiden ja hakukoneiden manipulointiin, eikä ihmisten tiedontarpeiden tyydyttämiseksi.
Ajankäyttö	Vaatii pitkän ajan investointia verkkosivuun. Sijoituksien paranemiseen kuluu aikaa.	Menetelmiä käyttämällä pyritään yleensä nopeisiin tuloksiin hakutuloksien paranemisessa.
Riskit	Menetelmien käyttöön ei liity riskejä.	Jos hakukone havaitsee menetelmien käytön, sivu voidaan esimerkiksi poistaa hakutuloksista.
Luotettavuus	Vaikka tulokset ovat hitaita, ne ovat kestäviä.	Johtuen sääntöjen vastaisuudesta verkkosivu voidaan poistaa milloin tahansa hakutuloksista.

Taulukko 1: Valkohattuisen ja mustahattuisen hakukoneoptimoinnin keskeisiä eroja (Varsha et al., 2021)

3.5 Harmaahattuinen hakukoneoptimointi

Rajanveto sallitun ja ei-sallitun hakukoneoptimoinnin välillä voi olla subjektiivinen päätös. Hakukoneoptimointia, joka ei ole selkeästi valko- tai mustahattuista, voidaan pitää *harmaahattuisena hakukoneoptimointina* (grey hat tai sham). Toisin kuin mustahattumenetelmiä hyödyntävä sivu, harmaahattuista hakukoneoptimointia hyödyntävä sivu voi sisältää informaatiotarpeen tyydyttävää sisältöä. Se kuitenkin sisältää lisäksi ainoastaan hakukoneoptimointitarkoituksessa luotua sisältöä, kuten avainsanoja, joita on huomattavan runsas määrä. Harmaahattuista hakukoneoptimointia hyödyntävä sivu voi sijoittua hakutuloksissa korkeammalle kuin se ansaitsisi sivun sisällön relevanssiin nähden. (Raiber et al., 2013)

Sivu voidaan luokitella harmaahattuista hakukoneoptimointia hyödyntäväksi, jos se täyttää jonkin seuraavista ehdoista:

1. Sivu sisältää tekstisisältöä, joka ei täytä mitään informaatiotarvetta.
2. Sivu sisältää tekstisisältöä, joka vaikuttaa liittyvän informaatiotarpeeseen, jonka jokin toinen sivun osio täytti, mutta sisältää paljon keinotekoisia, toistuvia tai muuten tarpeettomia lisäsanoja tai lauseita, jotka ovat lisätty sivulle vain hakutulossijoituksen nostamiseen.
3. Sivulla on osioita, joiden sisältö on kopioitu toiselta sivulta (kuten Wikipediasta) tarkoituksenaan oletettavasti vain nostaa sivun hakutulossijoitusta.

(Raiber et al., 2013)

4 Mustahattuhakukoneoptimointi yleisesti

Tässä luvussa käsitellään yleisesti mustahattuhakukoneoptimointia, sen vastatoimia ja negatiivista hakukoneoptimointia. Alaluvussa 4.1 kuvaillaan mustahattuhakukoneoptimointiin liittyviä käsitteitä ja kuvaillaan mustahattuhakukoneoptimoinnin yleisiä piirteitä, vaikutuksia, menetelmien luokittelua sekä motiivia. Alaluvussa 4.2 kuvaillaan mustahattuhakukoneoptimoinnin vastatoimia, niiden motiivia ja sitä, mikä aiheuttaa vastatoimia. Alaluvussa 4.3 kuvaillaan negatiivista hakukoneoptimointia ja injektioita.

4.1 Mustahattumenetelmät yleisesti

Spämmi

Mustahattuhakukoneoptimoinnin tutkimuksen yhteydessä kirjallisuudessa käytetään termiä spämmi. Spämmi on dokumentin sisältämää harhaanjohtavaa, epäsoveliasta ja epärelevanttia informaatiota, jota hyödyntäen pyritään saavuttamaan liiketoiminnallista hyötyä (Croft et al., 2009). Kirjallisuudessa käytetään hakukoneisiin kohdistuvaan spämmiin viitattaessa eri termejä. Croft ja muut (2009) sekä Shahzad ja muut (2021) käyttävät termiä *spamdexing,* kun taas Somani ja Suman (2011) käyttävät termiä *spamming*. Kumar ja muut (2016) käyttävät termejä *search engine spam* ja *web spam*. Viitattaessa mustahattuhaisen hakukoneoptimoinnin harjoittamiseen Liu ja muut (2020) käyttävät termiä *web spamming,* kun taas mustahattuista hakukoneoptimointia hyödyntävään verkkosivuun he viittaavat termillä *web spam.*

Mustahattuhakukoneoptimoinnin piirteet

Mustahattuisessa hakukoneoptimoinnissa pyritään nostamaan verkkosivun tai -sivuston hakutulossijoitusta (Kumar et al., 2016). Mustahattumenetelmille on ominaista niitä hyödyntämällä saavutettu perusteettoman korkeaksi tulkittavissa oleva hakutulossijoitus (Somani & Suman, 2011). Perusteettomuus johtuu verkkosivun todellisesta alhaisesta arvosta tiedonhakijalle (Li, 2014). Mustahattumenetelmin verkkosivu pyritään tarkoituk-

sellisesti saamaan vaikuttamaan todellista relevantimmalta (Shahzad et al., 2021). Verkkosivu, joka saa hakukonealgoritmin perusteetta viittaamaan itseensä tai toiseen sivuun, on spämmiä (Ghiam, 2012).

Mustahattuinen hakukoneoptimointi voi saada aikaan tiedonhakijan kannalta epäluotettavia ja epämiellyttäviä hakutuloksia (Somani & Suman, 2011). Mustahattumenetelmin ei pyritä luomaan verkkosivulla vierailevalle selaimen käyttäjälle laadukasta ja informatiivista sisältöä, vaan nostamaan hakukoneiden ryömijät ja algoritmit huomioon ottaen sivun hakutulossijoitusta (Jha & Saraswat, 2018).

Mustahattumenetelmin ei pyritä pelkästään optimoimaan verkkosivua, vaan manipuloimaan hakutuloksia (Fox, 2008). Mustahattuisen hakukoneoptimoinnin perusideana on pyrkiä johtamaan hakukoneiden järjestelyalgoritmeja harhaan (Kumar et al., 2016). Algoritmeja pyritään harhauttamaan esimerkiksi jäljentämällä sisältöä tai hyödyntämällä liikaa avainsanoja (Varsha et al., 2021). Mustahattuinen hakukoneoptimointi on tarkoituksellista manipulaatiota verkkosivun hakutulossijoituksen nostamiseksi (Li, 2014).

Siinä missä valkohattuinen hakukoneoptimointi tuo esille relevantteja hakutuloksia, mustahattuinen hakukoneoptimointi saa irrelevantin ja huonolaatuisen sisällön sijoittumaan hakutuloksissa laadukasta ja hyödyllistä sisältöä korkeammalle. Siksi mustahattuinen hakukoneoptimointi on haitallista hakutuloksille. Hakukoneoptimointia voidaan siis pitää mustahattuisena, jos sen tarkoituksena on vain järjestelyalgoritmeihin vaikuttaminen, eikä niinkään verkkosivun sisällön parantaminen. (Fox, 2008)

Mustahattuhakukoneoptimoinnin vaikutukset

Mustahattuinen hakukoneoptimointi luo rasitusta hakukoneille. Sen seurauksena hakukoneille aiheutuu enemmän sisältöä indeksoitavaksi (Li, 2014). Hakukone ryömii ja indeksoi jokaisen hakutuloksissaan esittämän hakutuloksen (Svore et al., 2007). Hakukone prosessoi kaikki ryömimänsä verkkosivut ja suorittaa vertailun jokaisen haun yhteydessä. Tämä kaikki kuluttaa kaistanleveyttä, tallennustilaa sekä suorittimien toimintaa. (Ntoulas et al., 2006) Mustahattuinen hakukoneoptimointi kuluttaa hakukoneiden resursseja (Svore et al., 2007) ja heikentää niiden järjestelymekanismeja (Li, 2014).

Mustahattuinen hakukoneoptimointi tuottaa tiedonhakijan intressien kannalta heikkoja hakutuloksia (Li, 2014). Se edistää hakijan tiedontarvetta vastaamattoman (Wang et al., 2014), huonolaatuisen ja haitallisen sisällön tulemista tiedonhakijan saataville (Li, 2014). Sen seurauksena tiedonhakija voi esimerkiksi ohjautua haittaohjelmia, tietojen kalastelua, tai väärennettyjen tuotteiden kauppaa sisältäville sivuille (Wang et al., 2014). Huonolaatuiset hakutulokset voivat kannustaa tiedonhakijaa vaihtamaan hakukonetta (Svore et al., 2007).

Mustahattuinen hakukoneoptimointi vääristää hakutuloksia (Wang et al., 2014)**.** Hyvät hakutulossijoitukset voivat olla verkkosivulle taloudellisesti tuottoisia. Verkkosivut

saattavat kokea mustahattumenetelmien käytön kannustavana, jos mustahattumenetelmiä hyödyntävät verkkosivut sijoittuvat hakutuloksien kärkeen. (Svore et al., 2007)

Mustahattumenetelmien luokittelu

Mustahattumenetelmät voidaan luokitella sen perusteella, onko niiden tarkoitus tehostaa verkkosivun hakutulossijoittumisen nousua (boosting techniques), vai piilottaa tehostavien menetelmien käyttö (hiding techniques). Tehostavia menetelmiä hyödyntäen pyritään lisäämään verkkosivun kävijäliikennettä ja nostamaan verkkosivun hakutulossijoitusta. Tehostavat menetelmät voidaan edelleen jakaa sivun sisäisiin ja ulkoisiin mustahattumenetelmiin. (Gyongyi & Garcia-Molina, 2005; Somani & Suman, 2011)

Piilotusmenetelmien hyödyntämisen tarkoituksena on pyrkiä estämään hakukoneita havaitsemasta tehostavien menetelmien käyttö. Piilotusmenetelmin pyritään harhauttamaan selaimen käyttäjiä ja hakukoneiden ryömijöitä (Gyongyi & Garcia-Molina, 2005; Somani & Suman, 2011). Piilotusmenetelmin pyritään usein esittämään selaimen käyttäjälle ja hakukoneen ryömijälle eri sisältöä, tai saamaan osan sivun sisällöstä käytännöllisesti katsoen näkymättömäksi selaimen käyttäjälle. Tehostavien menetelmien piilottaminen voidaan toteuttaa esimerkiksi sivun sisältämän tekstin piilottamisella. Muita menetelmiä ovat *cloaking* ja *doorway*-sivut. (Malaga, 2008) Cloaking-menetelmää ja doorway-sivuja käsitellään luvussa 8.

Hakutulossijoituksen nostamisen tehostamiseksi optimoitu verkkosivu ei välttämättä ole käyttäjäystävällinen. Näin ollen selaimen käyttäjälle ja hakukoneen ryömijälle voidaan esittää eri sisältö. Selaimen käyttäjälle esitetään käyttäjäystävällinen verkkosivu, kun ryömijälle esitetään optimoitu sisältö ilman sivun suunnitteluelementtejä. (Malaga, 2008) Verkkosivun sisältämän spämmin piilottamisen voi toteuttaa sivun HTML-dokumentin tai CSS-tyylisivun tyylillisillä muutoksilla, tai käyttämällä sivun visuaalisen elementin piilottavaa skriptiä, esimerkiksi asettamalla näkyvän HTML-tyyliattribuutin arvoksi false. (Gyongyi & Garcia-Molina, 2005)

Edellä esitetyn mustahattumenetelmien luokittelun lisäksi Shahzad ja muut (2021) jakavat menetelmät kolmeen luokkaan, jotka ovat *sisältöperustaiset menetelmät* (content-based spamdexing), *linkkiperustaiset menetelmät* (link-based spamdexing) ja cloaking. Ghiam (2012) jakaa menetelmät niin ikään kolmeen luokkaan, jotka ovat linkkiperustaiset menetelmät, sisältöperustaiset menetelmät ja piilotusmenetelmät, joista viimeksi mainitun hän jakaa cloaking- ja *uudelleenohjausmenetelmiin*. Piilotusmenetelmät, kuten cloaking, ovat sisältö- ja linkkiperustaisten menetelmiin verrattuna haastavampia havainnoida. Cloaking -menetelmän havainnointi vaatii useiden sivun versioiden vertailua. (Ghiam, 2012)

Mustahattuhakukoneoptimoinnin motiivi

Mustahattuinen hakukoneoptimointi mahdollistaa hakutulossijoitusten nopean kasvun (Jha & Saraswat, 2018). Hakutulossijoitusten kasvattamisella pyritään lisäämään verkkosivun kävijäliikennettä (Ghiam, 2012). Kävijäliikenteen kasvu voi olla verkkosivun omistajalle taloudellisesti tuottoisaa (Jha & Saraswat, 2018; Ghiam, 2012). Verkkosivu voi toimia esimerkiksi yrityksen ja sen tuotteiden esittelyn välineenä (Ghiam, 2012). Esimerkiksi väärennettyjä tuotteita myyvät verkkokaupat voivat saavuttaa kuukausien aikana satoja tuhansia tilauksia korkeiden hakutulossijoituksien myötä (Wang et al., 2014). Mustahattuisen hakukoneoptimoinnin harjoittajat voivat ansaita huomattavia tuloja esimerkiksi luomalla automaattisesti tuhansia korkealle sijoittuvia sivustoja. Vaikka sivustoja hyödyntäen ansaittu rahasumma on vain muutamia senttejä sivustoa kohden päivässä ja sijoitukset laskevat nopeasti, sivustojen suuri määrä tekee toiminnasta tuottavaa. (Malaga, 2008) Kävijäliikenteen kasvattamisella voidaan pyrkiä myös esimerkiksi haittaohjelmien levittämiseen (Ghiam, 2012).

Mustahattuinen hakukoneoptimointi vaatii valkohattuista vähemmän panostusta hakutulossijoitusten kasvattamiseksi (Jha & Saraswat, 2018). Valkohattuinen hakukoneoptimointi vaatii enemmän työtä ja sijoituksien nousu vie enemmän aikaa (Ma, 2018). Valkohattuinen hakukoneoptimointi vaatii enemmän rahaa ja muita resursseja, mikä on aiheuttanut mustahattuisten menetelmien suosion kasvun (Kumar et al., 2016). Mustahattumenetelmin saavutettu hakutulossijoitus ei kuitenkaan ole yhtä kestävä kuin valkohattumenetelmin saavutettu sijoitus. Mustahattumenetelmien käyttöön liittyy riski tulla poistetuksi hakuindeksistä tai alennetuksi hakutuloksissa. (Jha & Saraswat, 2018)

Hakukoneoptimoinnin harjoittajan on otettava huomioon käyttämiensä menetelmien seuraukset ja pohdittava, mihin hän optimoinnilla pyrkii. Tehdessään valintaa menetelmien välillä optimoinnin harjoittajan tulisi harkita sivun arvoa, saavutetun hyödyn kestoa sekä optimointiin liittyviä riskejä. Musta- ja valkohattumenetelmät tuovat sivustolle erilaista arvoa ja niiden avulla saavutettu hyöty on kestoltaan erilaista. (Jha & Saraswat, 2018)

Mustahattumenetelmiä hyödyntävät sivustot ovat usein yhteydessä erilaisiin huijauksiin. Sivustot voivat esimerkiksi sisältää haittaohjelmia tai tietojen kalastelua, olla yhteydessä *bottiverkkoon* (botnet), tai ne voivat olla hakkeroituja. (Pevtsov & Volkov, 2013) Mustahattumenetelmiä hyödyntävien sivustojen sisältö voi käsitellä esimerkiksi uhkapelejä tai lääkealaa (Yang et al., 2020).

4.2 Mustahattuhakukoneoptimoinnin vastatoimet yleisesti

Vastatoimien motiivi

Tässä tutkielmassa mustahattuhakukoneoptimoinnin vastatoimilla viitataan menetelmiin, joilla pyritään vähentämään mustahattuhakukoneoptimoinnin vaikutuksia. Tällaisia menetelmiä ovat rangaistusmenetelmät sekä keinot mustahattumenetelmiä hyödyntävien verkkosivujen havainnoimiseksi.

Mustahattuinen hakukoneoptimointi voi merkittävästi heikentää hakutulosten laatua, joten hakukoneyhtiöiden tulee kyetä havainnoimaan ja poistamaan mustahattumenetelmiä harjoittavat sivut hakuindekseistään (Croft et al., 2009). Hakukoneyhtiöt pyrkivät minimoimaan mustahattuisen hakukoneoptimoinnin vaikutuksia (Li, 2014), ja ne voivat antaa verkkosivulle rangaistuksen havaitessaan mustahattumenetelmien käytön (Lynn et al., 2015). Mustahattumenetelmien käytön havaitessaan hakukoneyhtiöt voivat laskea verkkosivun sijoitusta äkisti tai poistaa sen hakutuloksista kokonaan (Jha & Saraswat, 2018; Fox, 2008; Duk et al., 2013; Ma 2018; Wang et al., 2014). Siksi mustahattumenetelmiä hyödyntävät sivut ovat usein lyhytikäisiä (Duk et al., 2013; Ma 2018). Rangaistukset voivat olla kestoltaan esimerkiksi tuntien tai kuukausien mittaisia ja niillä voi olla verkkosivua ylläpitävälle taholle taloudellisia vaikutuksia (Lynn et al., 2015).

Vastatoimet

Hakukoneyhtiöt voivat antaa rangaistuksen automaattisesti tai manuaalisesti. Hakukoneen algoritmin antaessa automaattisen rangaistuksen manuaalisen rangaistuksen toteuttaa ihminen. Manuaalisesti rangaistavan verkkosivun ylläpito voi saada hakukoneyhtiöltä varoituksen ja parannusehdotuksen. Hakukone voi hakutulossivullaan esittää hakutuloksen yhteydessä varoituksen hakukoneen käyttäjälle (Lynn et al., 2015; Wang et al., 2014) Varoitus voidaan esittää myös välilehdellä, joka ilmestyy käyttäjän valitessaan hakutuloksen (Lynn et al., 2015).

Hakukoneyhtiöt ovat kehittäneet hakukonealgoritmejaan vähentääkseen mustahattuisen hakukoneoptimoinnin vaikutuksia. Esimerkiksi Google pyrkii Panda-algoritmiaan hyödyntäen kohottamaan sisältöään säännöllisesti ja kattavasti päivittävien verkkosivujen sijoitusta poistamalla huonolaatuista sisältöä hakutuloksistaan. Lisäksi hakutuloksia pyritään kehittämään linkkejä tutkivaa Penguin-algoritmia ja hakujen kontekstuaalisuutta käsittelevää Hummingbird-algoritmia hyödyntäen. Varsha et al., 2021) Googlen algoritmeja käsitellään luvussa 9.6. Sivusto voidaan ajaa alas oikeusteitse, mikä on hidas ja tehoton prosessi. Alas ajetun sivuston ylläpitäjä voi luoda toiminnalleen uuden verkkotunnuksen. (Wang et al., 2014)

Vastatoimien aiheuttajat

Vastatoimia voi aiheuttaa

1. Toistuva hakukoneyhtiöiden asettamien laatumääräysten vastaisuus
2. Spämmiksi tulkittava sisältö esimerkiksi sivun foorumissa tai kommenttiosiossa
3. Huomattava määrä sivulle osoitettuja epäluonnollisia, keinotekoisia, petollisia tai manipuloivia linkkejä
4. Sellaisen *verkkoisännöintipalvelun* (hosting service) käyttö, jossa huomattava osa muista sivustoista sisältää spämmiä
5. Haittaohjelmat
6. Hakkeroitu sivu

Hakukoneyhtiöt saattavat poistaa edellä kuvattuja ominaisuuksia sisältävät sivut tai sivustot hakuindeksistään, tai alentaa niiden hakutulossijoitusta. (Lynn et al., 2015)

Kilpajuoksu vastatoimien ja mustahattumenetelmien välillä

Hakukoneiden järjestelyalgoritmien ominaisuudet ovat manipuloitavissa ja järjestelyalgoritmien kehittyessä myös manipulointimenetelmät kehittyvät (Pevtsov & Volkov, 2013). Ennen järjestelyalgoritmit järjestelivät hakutulokset ennen kaikkea sisältöperustaisesti ja tällöin manipulointikin keskittyi verkkosivujen sisällöllisiin ominaisuuksiin. Spämmillä pyrittiin vaikuttamaan esimerkiksi sivujen TF-IDF-arvoon esimerkiksi luomalla merkityksetöntä sisältöä optimoidulla avainsanatiheydellä. Kun linkkien merkitys hakutuloksien järjestelyperusteena kasvoi, myös manipulointikeinot keskittyivät enemmän linkkeihin. PageRank-arvoon pyrittiin vaikuttamaan esimerkiksi luomalla keinotekoisia linkkiverkostoja, kuten *linkkifarmeja* (luku 7.3) sekä luomalla maksettuja linkkejä. Koska hakukoneyhtiöt ovat sittemmin reagoineet sisältöön, linkkeihin ja painalluksiin perustuviin manipulointimenetelmiin, mustahattuisen hakukoneoptimoinnin huomio on siirtynyt verkkosivujen dynaamiseen sisältöön, kuten JavaScriptiin. Kun verkkosivun indeksointi perustuu sen HTML-sisältöön, hakukoneet eivät kerää informaatiota sivun sisältämien skriptien vaikutuksesta. Tästä syystä hakukoneyhtiöiden tulisi kyetä havainnoimaan JavaScriptin käyttöön perustuvaa manipulaatiota. (Pevtsov & Volkov, 2013) Vaikka hakukoneyhtiöt ovat käyttäneet useita menetelmiä mustahattuisen hakukoneoptimoinnin vastaiseen toimintaan, manipulointimenetelmien ja niitä hyödyntävien verkkosivujen kasvava määrä tekee niiden ehkäisemisestä haastavaa. Näin ollen kyky havainnoida manipulointimenetelmiä tarkasti ja tehokkaasti on tärkeää. (Liu et al., 2020)

4.3 Negatiivinen hakukoneoptimointi ja injektiot

Mustahattuisesta hakukoneoptimoinnista voi aiheutua verkkosivulle haittavaikutuksia hakukoneyhtiöiden antamien rangaistusten muodossa. Näitä haittavaikutuksia voidaan käyttää hyväksi verkkosivujen hakutulossijoituksien tahalliseksi alentamiseksi. *Negatiivisen hakukoneoptimoinnin* harjoittajat pyrkivät tahallisesti aiheuttamaan verkkosivulle mustahattuisesta hakukoneoptimoinnista aiheutuvia haittavaikutuksia. Negatiivisen hakukoneoptimoinnin menetelmät ovat samoja kuin mustahattuisen hakukoneoptimoinnin menetelmät, mutta niiden käyttötarkoitus on eri. Negatiiviseen hakukoneoptimoinnin tahot ovat hyökkääjä, kohdesivusto (joka on negatiivisen hakukoneoptimoinnin kohteena) ja hakukoneyhtiö. (Lynn et al., 2015) Negatiiviseen hakukoneoptimointiin on viitattu myös *keilaamisena* (bowling), mikä viittaa kilpailevan verkkosivun (negatiivisen hakukoneoptimoinnin kohdesivun) poistamiseen kilpailusta heikentämällä sen hakutulossijoitusta (Malaga, 2008).

Hyökkääjä voi pyrkiä tahallisesti häiritsemään kohdesivun normaalia toimintaa, aiheuttamaan sille rangaistuksen, tai muulla tavoin aiheuttamaan haittaa sivulle ja sen sidosryhmille. Hyökkääjä voi pyrkiä esimerkiksi hakkeroimaan kohdesivun, sijoittamaan sivulle spämmiä tai osoittamaan sivulle epäluonnollisia linkkejä. (Lynn et al., 2015) Hyökkääjä voi osoittaa kohdesivulle linkkejä sivustoilta, jotka hakukoneyhtiöt arvostavat alas madaltaen myös kohdesivun hakutulossijoitusta. Sijoitusta alentavia huonolaatuisia linkkejä osoittavia sivustoja voivat olla esimerkiksi linkkifarmit tai sisällöltään kyseenalaiset, esimerkiksi uhkapelejä käsittelevät sivustot. (Malaga, 2008)

Hyökkäyksen yhteydessä kohdesivulle voidaan *injektoida* sisältöä, eli sijoittaa luvatta hyökkääjän laatimaa sisältöä (Yang et al., 2020). Kohdesivulle voidaan esimerkiksi sijoittaa runsaasti avainsanoja aiheuttaen sivun hakutulossijoituksen heikkenemisen (Malaga, 2008). Injektio voidaan toteuttaa esimerkiksi HTML-injektiona (Malaga, 2008) tai SQL-injektiona (Yang et al., 2020). Tässä tutkielmassa injektion kohteena olevaan sivuun viitataan *haavoittuneena* sivuna.

Kohdesivun hakutulossijoituksen alentamisen lisäksi injektioita voidaan hyödyntää myös kyseenalaisen sisällön näkyvyyden lisäämiseksi. Sisältö voi liittyä esimerkiksi uhkapeleihin ja haavoittunutta sivua käytetään sisällön levittämisen välineenä. Haavoittuneelle sivulle lisätään mustahattuisia näkyvyyttä lisääviä ominaisuuksia, mikä lisää sisällön leviämistä. (Yang et al., 2020)

Verkkosivun korkean hakutulossijoituksen kannalta on tärkeää suojella sitä sivulle kohdistuvilta ulkoa tulevilta mustahattukampanjoilta. Oman verkkosivun kanssa samoja avainsanoja hyödyntävät ulkoiset sivustot voivat olla huonolaatuisia, mikä voi heikentää myös oman verkkosivun sijoittumista. Sen lisäksi omalle sivulle voi kohdistua sijoituksia alentavia linkkejä. Tällaiset kampanjat voivat heikentää oman sivun sijoittumista. (Jha & Saraswat, 2018)

5 Tutkimusmenetelmä

Tämän tutkielman tavoitteena on selvittää, minkälaisia mustahattuisia hakukoneoptimointimenetelmiä ja niihin kehitettyjä vastakeinoja tieteellisissä julkaisuissa on tutkittu. Tutkimusmenetelmänä on kirjallisuuskatsaus. Motiivina on ollut selvittää menetelmien kirjoa tieteellisen tutkimuksen pohjalta. Tutkielmassa suoritettiin systemaattinen haku neljään tietokantaan. Lähtökohtaisesti neljän tietokannan valinnan tarkoituksena on ollut aineiston riittävyyden varmistaminen (vähintään neljäkymmentä lähdettä). Systemaattisen tiedonhaun lisäksi suoritettiin tarpeen mukaan lisähakuja. Systemaattisessa tiedonhaussa käytetyt tietokannat ovat ACM Digital Library, Computer Science Database (ProQuest), IEEE Electronic Library, ScienceDirect (Elsevier). Lisähakujen yhteydessä käytetyt tietokannat ovat Andor ja Google Scholar.

Aineistojen valikoitumisen perusteena on niiden relevanssi tutkimuksen aiheeseen nähden. Relevantti aineisto käsittelee mustahattuisia hakukoneoptimointimenetelmiä tai niiden vastatoimia. Myös yleisesti hakukoneoptimointia käsittelevät aineistot katsottiin relevantiksi. Systemaattisessa tiedonhaussa hakuterminä käytettiin termiä “black hat”. Hakutermi valikoitui, jottei hakutulokset olisi rajautuneet liikaa. Jos hakuterminä olisi käytetty esimerkiksi termiä “keyword stuffing”, muita mustahattumenetelmiä käsittelevät aineistot eivät välttämättä olisi esiintyneet hakutuloksissa. Lisäksi hakutermeinä käytettiin ”SEO” tai ”search engine optimization”. Hakulausekkeena käytettiin:

```
”black hat” AND (”SEO” OR “search engine optimization”)
```

Kahteen ensimmäiseen tietokantaan kohdistuneissa hauissa käytettiin “SEO” -hakutermiä. Hakutermistä “SEO” kuitenkin luovuttiin sen tuottaessa epärelevantteja tuloksia, sillä “SEO” saattaa esiintyä muun sanan osana. Systemaattisen tiedonhaun vaiheessa haku suoritettiin järjestyksessä edellä mainittuihin neljään tietokantaan. Hakutuloksien relevanssin arvioimiseksi artikkeleista tutkittiin otsikko ja tiivistelmä sekä suoritettiin yleiskatsaus artikkeliin. Artikkeleihin tehtiin usein myös haku, josta kävi ilmi mustahattumenetelmiä käsittelevä kohta. “Black hat” saattoi viitata johonkin muuhun kuin mustahattuiseen hakukoneoptimointiin. Näin ollen aineiston sisältäessä ”black hat”-termin, se ei välttämättä ollut relevantti artikkeli. Taulukosta 2 käy ilmi kunkin tietokannan kohdalla haussa käytetyt hakuehdot, haun ajankohta, kaikkien hakutuloksien määrä sekä relevanttien hakutulosten määrä. Jokainen relevantiksi merkitty aineisto merkittiin relevantiksi vain kerran. Jos yhteen tietokantaan kohdistuneen haun yhteydessä relevantiksi merkitty artikkeli esiintyi myöhemmin toiseen tietokantaan kohdistuneessa haussa, sitä ei enää merkitty relevantiksi. Systemaattisessa tiedonhaussa relevantteja tuloksia oli yhteensä 82, joista analyysin kohteeksi valittiin 31 artikkelia. Lisähaut huomioiden tutkielmassa käytettyjä lähteitä on yhteensä 49.

Tietokanta	Computer Science Database (ProQuest)	ACM Digital Library	IEEE Electronic Library (IEL) (haku 1)	IEEE Electronic Library (IEL) (haku 2)	Science Direct (Elsevier)
Hakupäivät	29.8.2022	7.-9.9.2022	12.9.2022	12.9.2022	13.9.2022
Hakutermit	"black hat" AND ("seo" OR "search engine optimization")	"black hat" AND ("seo" OR "search engine optimization")	"black hat" AND "search engine optimization" *	"black hat" AND "search engine optimization"	"black hat" AND "search engine optimization"
Hakukentät	Kaikki kentät	Kaikki kentät	All metadata	Metadata and full text	Kaikki kentät
Rajaukset	Scholarly journals	-	-	Journals, conferences	-
Hakutulokset	32	43	8	55	32
Relevantit tulokset	13	17	8	34	10

Taulukko 2: Kirjallisuuskatsauksen systemaattinen haku

* "seo"-hakusanasta luovuttiin, sillä se voi viitata johonkin sanan osaan, joka ei liity aiheeseen

6 Sivun sisäiset mustahattumenetelmät

Tässä luvussa kuvaillaan sivun sisäisiä mustahattumenetelmiä ja niiden vastatoimia. Alaluvussa 6.1 kuvaillaan yleisesti sivun sisäisiä mustahattumenetelmiä. Alaluvussa 6.2 kuvaillaan avainsanojen liiallista käyttöä ja sen määritelmää, avainsanojen käytön kohdistumista sivun sisältöön sekä kolmen suuren hakukoneen reaktioita eri avainsanojen käyttöön. Alaluvussa 6.3 kuvaillaan sisällön piilottamista ja alaluvussa 6.4 kuvaillaan automaattisesti luotua sisältöä. Mustahattumenetelmien yhteydessä kuvaillaan myös niiden vastatoimia.

6.1 Sivun sisäiset mustahattumenetelmät yleisesti

Sivun sisäiselle mustahattuhakukoneoptimoinnille on ominaista, että verkkosivulle luodaan todellisesta asiasisällöstä poikkeavaa sisältöä, jota voidaan pitää huonolaatuisena (Kumar et al., 2016). Sivun sisäiseen mustahattuhakukoneoptimointiin liittyy spämmin hyödyntäminen. Spämmi voi kohdistua eri rakenteellisiin osiin, kuten title-elementtiin, runkoon, metatunnisteeseen, ankkuritekstiin ja URL-osoitteeseen (Shahzad et al., 2021). Spämmi kohdistuu yleensä sivun tekstiin, millä pyritään vaikuttamaan sivun TF-IDF-arvoon, jota hakukoneet käyttävät yhtenä järjestelyn perusteena. Sivun sisäinen spämmi voi

vähentää verkkosivun laatua ja alentaa sivun *käyttäjäkokemusta* (user experience) (Pevtsov & Volkov, 2013). Sivun sisäiset mustahattumenetelmät ovat hakukoneille kaikista vahingollisimpia mustahattumenetelmiä, sillä ne ovat suosituimpia (Ghiam, 2012).

6.2 Avainsanojen liiallinen käyttö

Avainsanojen liiallisen käytön määritelmä

Yksi sivun sisäinen mustahattumenetelmä on avainsanojen liiallinen käyttö (keyword stuffing), jossa verkkosivun hakutulossijoitusta pyritään nostamaan lisäämällä sivulle ylimääräisiä avainsanoja (Fox, 2008; Croft et al., 2009). Avainsanoja käytetään verkkosivun tekstissä tai metatunnisteessa siinä määrin, ettei se ole ihmiselle luettavaa, normaalia tekstiä (Weideman, 2009). Hakukoneyhtiöiden ohjesäännöt sallivat avainsanojen käytön. Kun avainsanojen käytön tarkoituksena on vain sivun hakutulossijoituksen nostaminen eikä sivun sisällön parantaminen, avainsanoja voidaan katsoa olevan liian paljon. (Fox, 2008) Esimerkiksi Google, Bing ja Yahoo! eivät hyväksy ohjesäännöissään liiallista hakusanojen käyttöä, ja voivat väärinkäytön yhteydessä rangaista verkkosivua poistamalla sen hakuindekseistään. Ohjesäännöissä ei kuitenkaan ilmoiteta tarkkaa rajaa liialliselle avainsanatiheydelle. Aihetta tutkivassa kirjallisuudessa on poikkeavia näkemyksiä avainsanojen optimaalisesta tiheydestä sivun sisältötekstissä. (Zuze & Weideman, 2013) Hakukoneoptimoinnin kannalta verkkosivun sisältämien avainsanojen määrä tulisi tarjota riittävä sato hakukoneille, mutta sen ei pitäisi ajaa runsaudellaan pois sivulla vierailevia käyttäjiä (Visser & Weideman, 2011). Avainsanatiheyden ollessa liian korkea hakukone voi tulkita verkkosivun sisältävän spämmiä, mutta myös liian matala avainsanatiheys voi heikentää verkkosivun hakutulossijoittumista (Zuze & Weideman, 2013).

Verkkosivun sanamäärä kasvaa liiallisen avainsanojen käytön myötä. Vaikkei sivun suuri sanamäärä sellaisenaan ole vahva indikaattori, suuren sanamäärän ja spämmin välillä on havaittu korrelaatio. Myös sisällön suurella toistuvuudella on havaittu olevan yhteys spämmiin, sillä toistamalla avainsanoja mustahattumenetelmien harjoittajat pyrkivät vaikuttamaan järjestelyalgoritmeihin. Spämmin aiheuttaessa runsaasti toistuvuutta sivun kompressoitavuutta voidaan hyödyntää spämmin havaitsemiseen. Kompressoimalla verkkosivu sen sisällöstä karsitaan epäolennainen sisältö ja sivun kokoa voidaan pienentää. (Ntoulas et al., 2006)

Avainsanojen liiallinen käyttö

Ylimääräisiä avainsanoja voi olla kymmeniä tai satoja (Ntoulas et al., 2006). Avainsanatiheyden lisäksi huomionarvoista on myös avainsanojen valinta (Pevtsov & Volkov, 2013). Liiallisen avainsanojen käytön yhteydessä avainsanat ovat aiheeltaan usein muuhun sivun sisältöön liittymättömiä (Ntoulas et al., 2006). Lisäämällä sivulle sisällön aiheeseen liittymättömiä avainsanoja verkkosivu pyritään saamaan vastaamaan laajem-

malle joukolle kyselyitä. Eri avainsanoja voidaan käyttää suuria määriä. (Gyongyi & Garcia-Molina, 2005). Näin sivun toivotaan keräävän enemmän vierailijoita (Ntoulas et al., 2006).

Verkkosivulle voidaan lisätä yleisesti verkkohauissa käytettyjä avainsanoja (Kumar et al., 2016), jotka liittyvät kyseisenä ajankohtana suosittuun aiheeseen (Wang et al., 2011). Aikaan sidotut suositut avainsanat ulottuvat laajemmalle hakijakunnalle, mutta niitä hyödyntävien sivujen korkea sijoitus on yleensä lyhytkestoista. Usein tällaisia avainsanoja hyödyntävien sivujen sijoitus heikkenee päivässä. (Wang et al., 2011) Avainsanat voivat myös liittyä pysyvämpään aiheeseen, kuten lääketeollisuuteen (Wang et al., 2011). Näin sijoitusta pyritään nostamaan tiettyjä kyselyitä kohtaan. Tekstin joukossa toistetaan tiettyjä termejä ja pyritään saamaan verkkosivulle korkea TF-IDF-arvo. (Gyongyi & Garcia-Molina, 2005). Pysyvämpiin aiheisiin liittyviä avainsanoja hyödyntävät sivut voivat pysyä saavuttamissaan sijoituksissaan kauemmin kuin ajankohtaan sidottuja avainsanoja hyödyntämällä, mutta eivät saavuta yhtä laajaa hakijakuntaa (Wang et al., 2011).

Avainsanoja voidaan hyödyntää myös *yhdistelmäsanoina* (composite words). Yhdistelmäsanoja voivat olla esimerkiksi “freepictures” tai “downloadvideo”, joita hyödyntämällä pyritään saamaan vastaavuus kyselyille, joissa ei ole käytetty sanavälejä. Sivun sisältämillä pitkillä sanoilla on havaittu olevan yhteys spämmiin. Ntoulasin ja muiden (2006) tutkimuksessa sivut, joiden keskimääräinen sanapituus oli kymmenen merkkiä, sisälsivät spämmiä. (Ntoulas et al., 2006)

Avainsanat sivun sisällössä

Huomionarvoista on avainsanatiheyden ja avainsanojen valinnan lisäksi myös avainsanojen sijoittelu (Pevtsov & Volkov, 2013). Avainsanoja voidaan lisätä esimerkiksi HTML- tai *PHP*-tiedostoihin. Kun kyseessä on HTML-sivu, kohteena ovat HTML-tunnisteet (Somani & Suman, 2011). Avainsanoja voi olla *kudottuna* **(**weaving**)** muualta kopioidun tavallisen sisällön seassa lauseisiin upotettuna (Gyongyi & Garcia-Molina, 2005).

Esimerkkejä sivun tekstikentistä, joihin spämmi voi kohdistua:

1. Sivun runko (Gyongyi & Garcia-Molina, 2005)
2. Title-elementti – Hakukoneiden järjestelyalgoritmit voivat antaa title-elementin sisällölle tärkeän aseman (Gyongyi & Garcia-Molina, 2005). Title-tunnisteen sisältämä suuri avainsanamäärä voi olla vahva viite spämmiin (Ntoulas et al., 2006).
3. Metatunniste – Metatunnisteeseen kohdistuvan spämmin suosiosta johtuen niiden arvo on vähentynyt hakukoneet järjestelyalgoritmeissa. Esimerkki metatunnisteeseen kohdistuvasta spämmistä:

   ```
   <meta name="keywords" content="buy, cheap cameras, lens, accessories, nikon, canon">
   ```

   (Gyongyi & Garcia-Molina, 2005)
4. Header-tunniste (Somani & Suman, 2011)
5. Heading-tunniste (Somani & Suman, 2011)
6. Ankkuriteksti - spämmi on kohdesivulle osoittavan linkin ankkuritekstissä. Esimerkki:

   ```
   <a href="target.html">free, great deals, cheap, inexpensive, cheap, inexpensive, cheap, free</a>
   ```

   (Gyongyi & Garcia-Molina, 2005)

   Spämmisivu voi sisältää pelkästään linkkejä, joiden ankkuritekstien on tarkoitus lisätä kohdesivujen näkyvyyttä (Ntoulas et al., 2006).
7. Verkkotunnuksen nimi – verkkotunnuksen nimi sisältää suositun kyselyn avainsanoja (Pevtsov & Volkov, 2013). Esimerkki URL-osoitteen sisältämistä avainsanoista:

   ```
   "buy-canon-rebel-20d-lens-case.camerasx.com"
   ```

   (Gyongyi & Garcia-Molina, 2005)

Avainsanatiheys ja kolme suurta hakukonetta

Zuze ja Weideman (2013) tutkivat rajanvetoa optimaalisen ja liiallisen avainsanatiheyden välillä. He käyttivät tutkimuksensa lähteinä tieteellistä kirjallisuutta, haastatteluja ja hakukoneyhtiöiden ohjesääntöjä. Lisäksi he tutkivat täyteavainsanojen käytön vaikutusta sivun hakutulossijoituksissa vertaillen kolmea eri hakukonetta, jotka ovat Google, Bing ja Yahoo!. Zuze ja Weideman (2013) tutkivat ensinnäkin, mikä on avainsanatiheyden vaikutus verkkosivun indeksoimiseen kuluvaan aikaan. Toiseksi he tutkivat, miten eri avainsanatiheydet vaikuttavat kyseisten hakukoneiden indeksointiprosessiin. Kolmanneksi he tutkivat, mikä avainsanatiheyden tason tulee olla, jotta hakukoneet tulkitsevat sen liialliseksi. He tutkivat myös, mikä on alan asiantuntijoiden määritelmä liialliselle avainsanojen käytölle. (Zuze & Weideman, 2013)

Zuze ja Weideman (2013) loivat viisi tietokoneaiheista sivustoa käyttäen avainsanaa "laptops". Sivustojen sisältö erosi ainoastaan avainsanatiheydessä. Sivustoilla käytettiin pelkästään HTML-tiedostoja, jotteivat hakukoneryömijät havaitsisi mustahattuista hakukoneoptimointia miltään muilta osin, kuin mahdollisesti avainsanatiheyden osalta. Esimerkiksi JavaScript ja Flash-tiedostoja ei käytetty. Verkkotunnukset olivat nimetty www.getlaptops1.co.za, www.getlaptops2.co.za ja niin edelleen. Sivustot ladattiin hakukoneiden saataville tunnin välein toisistaan. Tutkimuksessa viiteen sivustoon viitattiin lyhenteillä GLPS1, GLPS2 … GLPS5. Avainsanatiheydet olivat nousevassa järjestyksessä. GLPS1-sivustolla oli alhaisin avainsanatiheys (3,95%) ja GLPS5-sivustolla oli suurin avainsanatiheys (27,3%). Testi suoritettiin kahdessa osassa, jotka olivat samanlaisia, mutta avainsanatiheyksiä nostettiin toisessa osassa. Sivustojen GLPS1-5 avainsanatiheydet olivat keskinäisessä suhteessa samat testin kummassakin osassa. Toisessa osassa avainsanatiheydet vaihtelivat välillä 30,3-97,27%. (Zuze & Weideman, 2013)

Zuze ja Weideman (2013) pyysivät hakukoneoptimoinnin asiantuntijoilta arvioita siitä, miten sivustot indeksoituisivat. Asiantuntijoiden mukaan ensimmäiset kolme sivustoa alhaisimmilla avainsanatiheyksillä tulisivat indeksoiduiksi. Neljäs sivusto aiheutti mielipiteissä hajontaa, kun taas viidennestä sivustosta asiantuntijat olivat yksimielisempiä siitä, ettei sitä tultaisi indeksoimaan. Asiantuntijoiden näkemyksissä esiintyi hajontaa optimaalisen avainsanatiheyden rajoista. Kolmesta viiteen prosenttiin arvioitiin olevan optimaalinen avainsanatiheys. Sen alittava avainsanatiheys olisi riittämättömästi optimoitu, kun taas sen ylittävä olisi täyteavainsanojen käyttöä. Yksi asiantuntijoista arvioi kahdentoista prosentin ylittävän avainsanatiheyden olevan liikaa. (Zuze & Weideman, 2013).

Testin ensimmäisessä osassa sivustot ladattiin hakukoneiden indeksoitaviksi. Bing ja Yahoo!-hakukoneet indeksoivat kaikki viisi sivustoa. Googlen hakukone indeksoi neljä, ja jätti GLPS1-sivuston indeksoimatta. Google-hakukoneen indeksointitulos testin ensimmäisessä osassa oli ristiriidassa haastatteluista saatujen tulosten kanssa, sillä haastateltavat olivat arvioineet, että GLPS1 indeksoitaisiin ensimmäisenä, koska sillä katsottiin olevan luonnollisin avainsanatiheys. Tutkijat huomasivat, että kyseisen sivuston sisältö oli kopioitu eräälle cloaking-menetelmää hyödyntävälle sivustolle. Kyseinen cloaking-menetelmää hyödyntänyt sivusto oli Google-hakukoneen indeksissä, mutta tutkijoiden luoma sivusto ei indeksoitunut. Tutkijat arvelivat tämän olevan syynä siihen, ettei Googlen hakukone indeksoinut GLPS1-sivustoa. GLPS5 oli nopeimmin indeksoitu sivu, sillä Bing ja Yahoo!-hakukoneet indeksoivat sen enintään viidessä päivässä. Sivuston avainsanatiheys oli 27,3 prosenttia. Pisimmillään indeksointi kesti testin ensimmäisessä osassa 33 päivää. (Zuze & Weideman, 2013)

Testin toinen osa suoritettiin kuten ensimmäinen, mutta avainsanatiheyksiä nostettiin, kuitenkin pitäen suhteelliset erot ennallaan sivustojen välillä. Indeksointiajat vaihtelivat

yhdeksästätoista kahteenkymmeneenyhdeksään päivään. GLPS5 avainsanatiheys oli yli 97 prosenttia. Testin toisessa osassa Bing ja Yahoo!-hakukoneet indeksoivat jälleen kaikki sivustot. Googlen hakukone indeksoi ainoastaan GLPS2-sivuston, jonka avainsanatiheys oli 40 prosenttia. (Zuze & Weideman, 2013)

Testin ensimmäisessä osassa Googlen hakukone ei indeksoinut GLPS1-sivustoa, jonka avainsanatiheys oli 3,94 prosenttia. Sen sijaan Googlen hakukone indeksoi sivuston, jonka avainsanatiheys oli neljäkymmentä prosenttia. Zuze ja Weideman (2013) arvioivat, ettei testin toisen vaiheen ensimmäisen sivuston hylkääminen johtunut avainsanatiheydestä, vaan siitä, että testissä käytettyjen sivustojen sisältöä oli kopioitu muille sivustoille, mikä aiheutti sivuston hylkäämisen indeksistä. Tälle väitteelle ei kuitenkaan löydetty vahvistusta. Google oli ainoa hakukone, joka jätti osan sivustoista indeksoimatta. Google ei kuitenkaan ilmoittanut tutkijoille, että heidän sivustonsa on jätetty indeksin ulkopuolelle. (Zuze & Weideman, 2013)

Testin ensimmäisen osan indeksointiprosentti oli 93 ja toisen osan 73. Yhteensä indeksointiprosentti oli 83. Indeksointiajat olivat testin ensimmäisessä osassa neljästä päivästä kolmeenkymmeneenkolmeen päivään, kun taas toisessa osassa yhdeksästätoista kahteenkymmeneenyhdeksään päivään. Googlen hakukoneella kesti pidempään indeksoida sivut kuin Bing ja Yahoo! -hakukoneilla. Googlen hakukoneella kesti lyhimmillään 11 päivää sivuston indeksointiin, kun taas Bing ja Yahoo! -hakukoneilla kesti lyhimmillään korkeintaan viisi päivää. Haastateltavat arvioivat, että indeksointi kestäisi kolmesta päivästä kolmeen kuukauteen. Zuze ja Weideman (2013) tutkimuksen yhteenlaskettu keskimääräinen indeksointiaika on 18,9 päivää. (Zuze & Weideman, 2013)

Zuze ja Weideman (2013) esittävät tutkimuksensa pohjalta, että Bing ja Yahoo! -hakukoneiden tapauksessa verkkosivujen kehittäjät voivat sivua optimoidessaan hyödyntää runsasta avainsanatiheyttä. Myös verkkosivujen indeksointiaika osoittautui lyhyemmäksi kuin ennakkotietojen perusteella arvioitiin. Tutkimuksessa osoittautui myös, ettei hakukoneet reagoi suureen avainsanatiheyteen yhtä voimakkaasti kuin haastattelujen ja kirjallisuuskatsauksen perusteella saattoi olettaa. Zuze ja Weideman (2013) kuitenkin huomauttavat, että vaikkei runsas avainsanatiheys aiheuttaisikaan rangaistusta hakukoneen toimesta, se voi aiheuttaa sivulle huonon käytettävyyden. Sivulla vierailevalle kävijälle voi olla epämiellyttävää, että sivun sisällöstä suuri osa on avainsanoja. Tämä voi karkottaa kävijät sivustolta. (Zuze & Weideman, 2013)

Sopivasta avainsanatiheydestä ei ole yhtenäistä näkemystä, koska hakukoneyhtiöiden ohjesäännöt eivät anna tarkkaa ohjetta siitä, mikä avainsanatiheys on liikaa ja asiantuntijoilla sekä aihetta käsittelevä kirjallisuudella on toisistaan poikkeavia näkemyksiä sopivasta avainsanatiheydestä. Tämä aiheuttaa verkkosivuja optimoiville kehittäjille haasteita sopivan avainsanatiheyden löytämiseen. Verkkosivujen suunnittelijoiden on avainsano-

jen käytössään tähdättävä vaikeasti määriteltävään tasapainoon avainsanarikkaan sisältötekstin ja täyteavainsanojen käytön välillä. Vaikka Zuze ja Weideman (2013) suosittelevatkin avainsanojen käyttöä, he suosittelevat kehittäjiä tekemään verkkosivujen sisällöstä ennen kaikkea laadukasta. (Zuze & Weideman, 2013)

6.3 Sisällön piilottaminen

Usein avainsanoja piilotetaan liiallisen avainsanojen käytön yhteydessä. Piilotettu sisältö ei näy selaimen käyttäjälle ja se on tarkoitettu vain ryömijöiden havaittavaksi. Piilottaminen toteutetaan esimerkiksi käyttämällä sisältötekstissä näkymätöntä fonttia, tai sijoittamalla avainsanat elementteihin, jotka eivät näy sivulla (Fox, 2008). Tällaisia elementtejä ovat esimerkiksi kuvien alt-tekstit tai metatunnisteen sisältö. Avainsanat voivat sijaita myös HTML-kommenteissa. (Ntoulas et al., 2006) HTML-elementit voivat olla sivun taustan kanssa saman värisiä, tai niiden fontit voivat ovat huomaamattoman pieniä (Malaga, 2008). Esimerkki tekstin piilottamisesta HTML-dokumentissa:

```
<body background="white">
<font color="white">hidden text</font>
</body>
```

(Gyongyi & Garcia-Molina, 2005)

Piilotettavat elementit voidaan sisällyttää myös hyödyntämällä CSS hidden div-tunnisteita. Hakukoneyhtiöt ovat antaneet HTML-elementtien piilottamisesta sekä hidden div -tunnisteiden käytöstä sanktioita. Hidden div -tunnisteita voidaan kuitenkin käyttää myös hyvänlaatuisessa sivun suunnittelussa, joten sen käytöstä rankaiseminen ei ole ongelmatonta. (Malaga, 2008)

Visuaalisuutta korostavat sivun osat, kuten CSS, parantavat sivun luettavuutta. Koska mustahattuisessa hakukoneoptimoinnissa sisältö on usein tarkoitettu vain ryömijöiden tutkittavaksi, visuaalisuutta korostavan sisällön puute voi viitata spämmiin. (Ntoulas et al., 2006)

6.4 Automaattisesti luotu sisältö

Mustahattuisessa hakukoneoptimoinnissa pyritään usein tuottamaan sisältöä nopeasti useille sivuille, minkä tarkoituksena on kasvattaa sivujen kävijäliikennettä ja saada vierailijat esimerkiksi painamaan mainoksia. Sattumanvarainen sisältö voidaan luoda automaattisesti siihen soveltuvaa ohjelmistoa hyödyntäen. (Fox, 2008) Automaattisesti luotu sisältö voidaan täyttää suosituilla avainsanoilla (Somani & Suman, 2011). Automaattisesti luotu sisältö voi olla kieliopillisesti puutteellista, mitä voidaan hyödyntää spämmin tunnistamisessa (Ntoulas et al., 2006). Sivun sisältö voidaan myös kopioida toiselta si-

vulta (scraping) (Fox, 2008; Patil et al., 2021). Sisältö voidaan kopioida useista eri lähteistä. Eri sivuilta kopioidut lauseet tai virkkeet liitetään yhteen (phrase stitching) tekstin tuottamiseksi. (Gyongyi & Garcia-Molina, 2005)

Jotta hakukoneet eivät havaitsisi plagiointia, kopioitua sisältöä voidaan muokata hyödyntämällä siihen soveltuvaa ohjelmistoa. Kopioitava sisältö on usein peräisin suositulta verkkosivulta. Tekstiä pyritään muokkaamaan muuttamatta sen merkitystä korvaamalla sanoja tai lauseita synonyymein. Menetelmään viitataan termillä *spinning*, *text-spinning* (Shahid et al., 2017) tai *article spinning* (Jha & Saraswat, 2018).

Spinning voidaan toteuttaa automaattisesti tai manuaalisesti. Manuaalisesti se voidaan toteuttaa hyödyntämällä *crowdturfing-sivuja* tai muita halvan työvoiman palkkaamiseksi tarkoitettuja markkinapaikkoja (black hat marketplace). Automaattinen spinning voidaan toteuttaa hyödyntämällä siihen soveltuvaa ohjelmistoa, kuten The Best Spinner (TBS) tai Spinbot, jotka uudelleenjärjestävät virkkeitä ja korvaavat sanoja synonyymein. Ohjelmistoa käyttämällä yhdestä alkuperäisestä tekstisisällöstä voidaan luoda useita versioita. Ohjelmistojen tuottamat sisällöt eivät ole välttämättä yhtä luettavia kuin manuaaliset tuotetut, mutta automaattinen menetelmä on kustannustehokkaampi ja sen avulla dokumentteja voidaan luoda suuria määriä. (Shahid et al., 2017)

Spinning-menetelmä aloitetaan tyypillisesti syöttämällä valittu lähdedokumentti spinning-ohjelmistolle, joka luo siitä mahdollisesti useita versioita. Koska tulokset saattavat olla epäluettavia, ne voidaan tarkistaa automaattisella luettavuuden ja kieliopintarkistusohjelmistolla, joka on osassa spinning-ohjelmistoja sisäänrakennettuna. Epäluettavat versiot voidaan hylätä ja kelpuutetut dokumentit tarkistetaan plagioinnin havainnointityökalulla. Jos dokumentti läpäisee testin, se voidaan julkaista verkossa. (Shahid et al., 2017)

Hakukoneet tutkivat verkkosivujen sisältöä ja antavat sanktioita verkkosivuille ja verkkotunnuksille, joiden havaitaan sisältävän plagioitua sisältöä. Tyypillisesti hakukoneyhtiöt hyödyntävät verkkosivujen sisällön plagioinnin havainnointiin siihen tarkoitettuja automaattisia työkaluja. Plagioimisen tunnistamiseen on pääsääntöisesti kaksi tyyppiä. Ulkoisessa (extrinsic) havainnoinnissa tutkittavaa aineistoa verrataan referenssidokumenttikokoelman aineistoihin ja sisäisessä (intrinsic) havainnoinnissa tutkittavaa aineistoa arvioidaan erikseen. Hakukoneiden indekseissä on suuri määrä dokumentteja, joten hakukoneet voisivat teoriassa havainnoida plagiointia vertailua dokumenttien välillä. Tämä olisi kuitenkin haastavaa toteuttaa laajassa mittakaavassa, sillä indeksit ovat todella suuria, ja kasvavat jokaisen ryömityn dokumentin kohdalla. (Shahid et al., 2017)

Verkkosivun sisältävän spämmin havainnoimiseen on hyödynnetty esimerkiksi sähköpostisuodattimien kaltaisia kielimalleja ja luokittelijoita (Kumar et al., 2016). Shahid

ja muut (2017) loivat mallin spinning-ohjelmistoa hyödyntäen luodun sisällön tunnistamiseksi. Spinning-ohjelmiston tuottaman tekstin tunnistamiseen hyödynnettävien ominaisuuksien pääkategoriat ovat esitelty taulukossa 3.

Tekstin ominaisuuksien pääkategoriat
N-grammit, jotka mittaavat sanojen esiintyvyyttä
Tavanomaiset leksikaaliset ominaisuudet, esimerkiksi tavujen määrä kappaleessa
Sanaston rikkaus
Luettavuus
Syntaktiset ominaisuudet, eli merkkien ja sanojen syntaktiset ominaisuudet, esimerkiksi vokaalien ja konsonanttien järjestyksen vaihtelu
Sekavuus

Taulukko 3: Spinning-ohjelmiston tuottaman tekstin tunnistamiseen hyödynnettävien ominaisuuksien pääkategoriat (Shahid et al., 2017)

Jos taulukossa 3 esiteltyjä ominaisuuksia hyödyntäen todetaan, että teksti on luotu hyödyntäen spinning-ohjelmistoa, sen alkuperäinen vastine pyritään löytämään referenssidokumenttikokoelmassa. Ensin indeksikokoelmasta etsitään viisi tutkittavan dokumentin kanssa samankaltaisinta dokumenttia hyödyntämällä dokumenttien TF-IDF-arvoa. Vastaavuutta lasketaan myös sanajärjestyksen (word sequence alignment) perusteella. Jos sanajärjestyksen perusteella löydetään tietyn raja-arvon ylittävä vastaavuus, voidaan olettaa, että tutkittavan dokumentin lähdedokumentti on indeksissä. Shahidin ja muiden (2017) menetelmää voidaan hyödyntää spinning-ohjelmistolla luodun sisällön ja sen alkuperäisen lähteen tunnistamiseen ilman, että tunnetaan spinning-ohjelmiston hyödyntämä sanakirja. (Shahid et al., 2017)

7 Linkkiperustaiset mustahattumenetelmät

Tässä luvussa kuvaillaan linkkiperustaisia mustahattumenetelmiä ja niiden vastatoimia. Alaluvussa 7.1 kuvaillaan linkkiperustaisia mustahattumenetelmiä yleisesti. Alaluvussa 7.2 kuvaillaan linkkien osoittamista mustahattumenetelmänä. Alaluvussa 7.3 kuvaillaan linkkifarmeja ja niiden vastatoimia. Alaluvussa 7.4 kuvaillaan linkkien julkaisemista ja alaluvussa 7.5 blogien ja muiden sivujen ylläpitoa mustahattumenetelminä. Alaluvussa 7.6 kuvaillaan mustahattuhakukoneoptimoinnissa hyödynnettäviä injektioita. Alaluvussa 7.7 kuvaillaan linkkien ostamista, sekä kävijäliikennettä tarjoavia palveluja ja niiden vastatoimia.

7.1 Linkkiperustaiset mustahattumenetelmät yleisesti

Linkkiperustaisilla mustahattumenetelmillä pyritään nostamaan sivun hakutulossijoitusta vaikuttamalla niihin järjestelyalgoritmeihin, jotka järjestelevät hakutuloksia linkkiperustaisesti (Gyongyi & Garcia-Molina, 2005). Järjestelyperusteena on sivun kävijäliikenteen

lisäksi sivun osoittamat linkit ja sivulle osoitetut linkit sekä linkkien ominaisuudet (Somani & Suman, 2011). Hakutuloksien järjestelyn perusteena on sivulle osoittavien linkkien lukumäärä (link population) (Somani & Suman, 2011; Shahzad et al., 2021). Sivun ulkoiset, eli toiselta sivustolta osoitetut linkit lisäävät sivun arvoa (Fox, 2008). Sivulle osoitetut *luotettavat linkit* (trusted links) lisäävät sivun arvoa. Luotetut linkit ovat luotetuilta sivustoilta osoitettuja linkkejä (Somani & Suman, 2011).

Sivun laadukas sisältö voi vaikuttaa positiivisesti sivulle osoitettavien linkkien määrään etenkin, jos linkin osoittavan sivuston sisältö käsittelee samaa aihetta (Jha & Saraswat, 2018). Linkkien osoittavien verkkosivujen tulisi käsitellä samaa aihetta kuin sivu, jolle linkki on osoitettu (Killoran, 2013)

7.2 Linkkien osoittaminen kohdesivulle ja kohdesivulta muille sivuille

Sivun hakutulossijoitukseen voi pyrkiä vaikuttamaan osoittamalla sivulle linkkejä muilta sivustoilta. Sivun PageRank-arvoon voi pyrkiä vaikuttamaan kasvattamalla sivulle osoitettujen linkkien lukumäärää. (Gyongyi & Garcia-Molina, 2005) Sivulle osoittavien linkkien kerryttäminen keinotekoisesti vain hakutuloksien manipuloimiseksi on hakukoneyhtiöiden ohjesääntöjen vastaista (Fox, 2008). Linkkejä voi pyrkiä kerryttämään luotetuilta sivustoilta (Somani & Suman, 2011). PageRank-algoritmi nostaa sivun sijoitusta, kun sille on osoitettu linkkejä arvokkailta sivuilta (Shahzad et al., 2021). Luotetuilta sivustoilta osoitetut linkit voivat heikentää hakukoneiden kykyä havainnoida mustahattuista hakukoneoptimointia (Ghiam, 2012).

HITS-algoritmi järjestelee hakutuloksia perustuen sivujen hub- ja auktoriteettiarvoihin. Sivun hub-arvoa voi pyrkiä kasvattamaan osoittamalla sivulta linkkejä tunnetuille sivuille. Sivun auktoriteettiarvoa taas voi pyrkiä kasvattamaan osoittamalla sivulle linkkejä sivuilta, joilla on korkea hub-arvo. (Gyongyi & Garcia-Molina, 2005)

Sivun hakutulossijoitukseen voi pyrkiä vaikuttamaan osoittamalla sivulta linkkejä muille sivustoille (Gyongyi & Garcia-Molina, 2005). Kohdesivun osoittamia linkkejä voi lisätä manuaalisesti, tai esimerkiksi kopioimalla verkkosivujen osoitteita osoitekirjastoista (kuten DMOZ Open Directory tai Yahoo! Directory), jotka mahdollistavat sivustojen selaamisen aiheiden perusteella (Gyongyi & Garcia-Molina, 2005).

Sivulla olevia linkkejä voidaan pyrkiä piilottamaan esimerkiksi muuttamalla ankkuritekstin mahdollisimman näkymättömäksi. Ankkuriteksti voidaan korvata esimerkiksi ankkurikuvalla, joka voi kokonsa vuoksi on vaikeasti havaittavissa selaimen käyttäjälle. Toisaalta ankkuriteksti tai -kuva voidaan muuttaa sivun taustan kanssa saman väriskeksi. (Gyongyi & Garcia-Molina, 2005) Esimerkki piilotetusta linkistä HTML-dokumentissa:

```
<a href="target.html"> <img src="tinyimg.gif"></a>
```

(Gyongyi & Garcia-Molina, 2005)

7.3 Linkkifarmi

Järjestelyalgoritmeihin voi pyrkiä vaikuttamaan toteuttamalla linkkifarmi. Linkkifarmi on useiden sivustojen joukko, jonka tarkoituksena on osoittaa linkkejä halutuille sivuille. (Gyongyi & Garcia-Molina, 2005; Somani & Suman, 2011; Patil et al., 2021) Linkkifarmi voi olla mustahattuisen hakukoneoptimoinnin harjoittajan itsensä luoma, tai mustahattuisen hakukoneoptimoinnin harjoittaja voi osallistua jo olemassa olevaan linkkifarmiin ja linkkien vaihtoon, jolloin joukko eri sivustoja vaihtavat linkkejä keskenään (Gyongyi & Garcia-Molina, 2005). Linkkifarmien yhteydessä linkit eivät usein ole osoitettu luotetuilta sivustoilta. Tästä syystä hakukoneiden on helpompi havaita linkkifarmit, kuin mustahattumenetelmät, jotka perustuvat luotettuihin linkkeihin. Tästä huolimatta linkkifarmit ovat haitallisempia hakukoneille. (Ghiam, 2012)

Linkkifarmiin kuuluvat mustahattusivut osoittavat usein linkin sivustolle, joka osoittaa linkin takaisin. Tätä ominaisuutta on hyödynnetty linkkifarmien havainnointiin *Parental penalty* -menetelmässä. Wu ja Davison (2005) havainnoivat linkkifarmeja tutkimalla sivun osoittamien linkkien ja sivulle osoittavien linkkien yhtäläisyyttä. He luokittelivat sivun linkkifarmiin kuuluvaksi, jos se sisälsi vähintään kolme linkkiä, joista oli osoitettu linkki takaisin. Tutkimusta jatkettiin laajennusvaiheeseen (expansion step) etenemällä sivuille, jotka olivat osoittaneet linkin linkkifarmiin kuuluville sivuille. Jos laajennusvaiheessa tutkittu sivu osoitti kolme linkkiä edellisessä vaiheessa linkkifarmiin kuuluviksi luokitelluille sivuille, sivu luokiteltiin myös linkkifarmiin kuuluvaksi. Tätä menetelmää jatkettiin iteratiivisesti linkkifarmin täydentämiseksi. (Wu & Davison, 2005) Parental Penalty -menetelmä ei havaitse duplikaatteja ja se saattaa virheellisesti luokitella mustahattusivuja hyvälaatuisiksi (Ghiam, 2012).

MLSA-algoritmi (Multi Level Link Structure Analysis) on kehittyneempi versio Parental Penalty -algoritmista. Sen toiminta perustuu havaintoon, jonka mukaan linkkifarmiin kuuluvan verkkotunnuksen sivujen joukossa on usein ainakin yksi toiselle verkkotunnukselle osoittava linkki. MLSA-algoritmi huomioi linkkien vaihtoa verkkotunnuksen sisäisten ja ulkoisten sivujen välillä. Tutkimusta laajennetaan niin monelle verkkotunnuksen sivulle, kuin on tarpeen tutkinnan tarkkuudesta riippuen. Algoritmi laskee kaikki sivulle tulevien ja sivulta lähtevien linkkien yhteiset verkkotunnukset. Jos lukema ylittää tietyn raja-arvon, sivu merkitään linkkifarmiin kuuluvaksi. Vaikka MLSA on edistyneempi kuin Parental Penalty, se luokittelee virheellisesti hyvälaatuisia sivuja linkkifarmiin kuuluvaksi. Ghiamin (2012) mukaan MLSA -algoritmia voi kehittää lisäämällä siihen sivun sisällönarviointiominaisuuden. (Ghiam, 2012)

Verkkosivujen ja niiden osoittamien linkkien muodostama verkko voidaan mallintaa graafina, jossa sivut ovat *solmuja* (node) ja linkit *kaaria* (edge). Mustahattusivuja ja linkkifarmeja voidaan pyrkiä tunnistamaan tutkimalla graafin ominaisuuksia tiettyjä raja-ar-

voja asettamalla. Ghiam (2012) viittaa menetelmään termillä *Link farm properties*. Graafin tutkittavat ominaisuudet ovat *degree distribution* ja *average path length*. Degree distribution on todennäköisyys sille, että satunnaisesti valitulla solmulla on tietty määrä kaaria. Average path length taas mittaa koko verkon kokoa käyttämällä lyhimmän polun kokoa ja solmujen määrää. Ghiamin (2012) mukaan Link farm properties -menetelmä havaitsee mustahattusivuja tehokkaammin ja aiheuttaa vähemmän virhearvoja kuin Parental Penalty ja MLSA. (Ghiam, 2012)

7.4 Linkkien julkaiseminen muilla sivustoilla

Mustahattuisen hakukoneoptimoinnin harjoittaja voi osoittaa linkkejä haluamilleen verkkosivuille omistamaltaan verkkosivulta. Linkkejä voi osoittaa myös sivuilta, joihin mustahattuhakukoneoptimoinnin harjoittajalla on rajallinen hallinta, sillä osa verkkosivuista sallii esimerkiksi kommenttien kirjoittamisen. (Gyongyi & Garcia-Molina, 2005) Linkkejä voi lähettää esimerkiksi blogeihin ja foorumeille (Somani & Suman, 2011). *Spämmikommentoinnissa* (spam commenting) mustahattuhakukoneoptimoinnin harjoittaja lähettää kohdesivulle osoittavia linkkejä erilaisiin kommenttikenttiin. Kommenteilla ei ole informatiivista viestiä, eikä kohdesivu liity kommentoitavan sivuston aiheeseen. (Jha & Saraswat, 2018) Vaikka kommentoinnin sallivia sivuja moderoitaisiin manuaalisesti, linkkien havaitsemista voidaan häiritä linkin piilottamiseen toteuttavalla menetelmällä. (Gyongyi & Garcia-Molina, 2005) Linkkejä voidaan lähettää myös sosiaalisen median sivuille. Linkkejä voidaan naamioida myös sosiaalisen median palveluissa. Esimerkiksi Facebook-palveluun lähetetty linkki voi olla naamioitu sovellukseksi. (Somani & Suman, 2011) Linkki voidaan lähettää jollekin relevantille sivulle (honey pot), jonka toivotaan suosionsa vuoksi kasvattamaan kohdesivun kävijäliikennettä (Gyongyi & Garcia-Molina, 2005; Somani & Suman, 2011)

7.5 Blogien ja muiden sivujen ylläpito

Mustahattuisen hakukoneoptimoinnin harjoittaja voi luoda useita blogisivuja, jotka ohjaavat vierailijan kohdesivulle (spam blog). Blogisivujen linkkejä sijoitetaan muille sivustoille, kuten blogeihin tai foorumeille. Uudelleenohjaussivuina toimivat blogisivut löytyvät myös hakukoneiden hakutuloksista. (Somani & Suman, 2011) Mustahattuhakukoneptimoinnin harjoittaja voi ylläpitää useita blogisivustoja, luoda uusia blogisivuja ja lisätä niihin linkkejä. Blogisivustojen palvelimille lähetetään automaattinen viesti blogin päivittämisestä (ping) ja jatkuvat päivitykset houkuttelevat ryömijöitä. Menetelmään viitataan termillä *blog-ping.* (Malaga, 2008)

Mustahattuhakukoneoptimoinnin harjoittaja voi hyödyntää sivulleen kohdistuvaa spämmiä. *Blog-spam*-menetelmässä (jota ei pidä sekoittaa spam blog -menetelmään) hyödynnetään sivun *sisällönhallintajärjestelmää* (content management system) *spämmisuo-*

jan (spam protection) ollessa aktivoituna. Kun kyseiseen sivuun kohdistuu spämmäämisen tarkoituksessa lisätty linkki, siitä erotellaan avainsanat, poistetaan alkuperäinen linkki ja asetetaan oma linkki. Kohdesivustolle kohdistuvaa kävijäliikennettä voidaan lisätä käyttämällä nofollow-tunnistetta, joka lisätään HTML-elementtiin. Se mahdollistaa selaimen käyttäjän seuraavan linkkiä ilman, että hakukoneen ryömijä seuraa linkkiä. (Somani & Suman, 2011)

Mustahattuisen hakukoneoptimoinnin harjoittaja voi lisätä kohdesivuston arvoa ja kohdesivustolle osoittavia linkkejä ostamalla vanhentuneita verkkotunnuksia ja hyödyntämällä niiden mukanaan tuomia sivuja linkkien osoittamiseen (Gyongyi & Garcia-Molina, 2005). Halpoja verkkotunnuksia voi hyödyntää monimutkaisten sivustorakenteiden rakentamiseen (spider-pool), joilla hakukoneiden ryömijät joutuvat rakenteesta johtuen ryömimään tavallista enemmän (Yang et al., 2020).

7.6 Injektiot

Mustahattuhakukoneoptimoinnin harjoittaja voi saada luvattomasti pääsyn toisen tahon hallussa olevan sivuston muokkaamiseen ja käyttää hyväkseen sivuston arvoa nostaessaan kohdesivun hakutulossijoitusta. Tunkeutumisen kohteena oleva sivusto voi olla esimerkiksi osoitekirjasto, jolle mustahattuhakukoneoptimoinnin harjoittaja voi asettaa kohdesivulle osoittavia linkkejä. (Gyongyi & Garcia-Molina, 2005)

Tunkeutuja voi suorittaa kohteena olevalle sivustolle HTML-injektion. Tunkeutuja voi esimerkiksi sijoittaa kohdesivun linkin toisella sivustolla toimivaan hakuohjelmistoon. WebGlimpse on akateemisilla sivustoilla käytetty hakuohjelmistosivu. Esimerkiksi The Stanford Encyclopedia of Philosophy –sivusto on käyttänyt WebGlimpse-pakettia. Tunkeutuja voi valita seuraavan osoitteen ja korvata kohdan ##site## kohdesivuston URL-osoitteella ja kohdan ##word## valitsemallaan ankkuritekstillä:

```
http://plato.stanford.edu/cgi-
bin/webglimpse.cgi?nonascii=on&query=%22%3E%3Ca+href%3Dhttp%3A%2F%2
F##site##%3E##word##%3C%2Fa%3E&rankby=DEFAULT&er-
rors=0&maxfiles=50&maxlines=30&maxchar s=10000&ID=1
```

(Malaga, 2008)

Osa luotetuista sivustoista palauttavat *URL-kyselyn* (URL query string) hakutuloksissaan, vaikka muuta merkityksellistä sisältöä ei palautettaisikaan. Tätä ominaisuutta hyödynnetään *site free-ride* -menetelmässä, jossa luotetun sivuston arvoa hyödynnetään halutun sisällön sijoittumiseksi korkealle hakutuloksissa. Mustahattuhakukoneoptimoinnin harjoittaja voi luoda hakuun tarkoitetun URL:n (a search URL), joka viittaa jollekin luotetulle sivustolle sijoittaen kyselyyn (query string) halutun viestin, ja injektoida tämän valitsemalleen sivulle. (Du et al., 2016)

Hyötyäkseen haavoittuneista sivuista injektion toteuttajat pyrkivät tuottamaan sivuille kävijäliikennettä. Näin ollen injektiokampanjoissa haavoittuneita sivuja voidaan havainnoida tutkimalla sivuilla mahdollisesti esiintyviä muita mustahattumenetelmiä, kuten sivun avainsanoja, TF-IDF-arvoa, linkkejä ja HTML-rakennetta. Avainsanoja, joita on havaittu mustahattuhakukoneoptimoinnin yhteydessä, voidaan hyödyntää käyttämällä niitä verkkohauissa ja hyödyntämällä hakukoneen *related search* -ominaisuutta. Tutkittaessa sivun linkkejä on syytä kiinnittää huomio linkkien osoittamien sivustojen laatuun, sillä haitalliselle sivustolle johtava linkki on saatettu sijoittaa haavoittuneelle sivulle kävijäliikenteen kasvattamiseksi. Sivun CSS- ja JavaScript-tiedostot saattavat sisältää ominaisuuden, jonka tarkoitus on esimerkiksi piilottaa linkki:

```
<div style=”display:none;”>
<a href=http://eve.xyz/>mark six</a>
</div>
```

(Yang et al., 2020)

7.7 Linkkien ostaminen ja kävijäliikennettä tarjoavat palvelut

Yleistä

Mustahattuhakukoneoptimoinnin harjoittaja voi ostaa kohdesivulle osoittavia linkkejä (Fox, 2008). Linkkejä ostaessa kohdesivun sisällöllisellä laadulla ei ole merkitystä. Näin ollen linkkejä ei tarvitse ansaita laadukkaalla sisällöllä. Linkkejä ostaessa myös ankkuritekstin voi valita (Jha & Saraswat, 2018). Yksi suosituimmista Yandex-hakukoneeseen kohdistuvista linkkiperustaisista mustahattumenetelmistä on ollut linkkien ostaminen *linkkien välittäjältä* (link-broker). Hakukoneyhtiön hitaan reagoinnin vuoksi menetelmä on ollut pitkään suosittu ja tehokas tapa nostaa sivujen hakutulossijoitusta. (Pevtsov & Volkov, 2013)

Kävijäliikenteen kasvattaminen nostaa sivun hakutulossijoitusta. Kävijäliikennettä voidaan manuaalisesti kasvattaa hyödyntämällä yhteisöjä, joissa ihmiset tekevät ohjeiden mukaisia hakuja ja painavat tuloksia *pay-per-click*-mallin mukaisesti. Vastaava menetelmä voidaan suorittaa automatisoidusti bottiverkkoja hyödyntämällä. Bottiverkkojen käyttö on yksi c*lick fraud* -menetelmän muoto*,* joka on Pevtsovin ja Volkovin (2013) mukaan yksi haitallisimmista mustahattumenetelmistä, jolla on huomattavia vaikutuksia koko internetin ekosysteemiin. Yksi hakukoneyhtiöiden tärkeimmistä tehtävistä on reagoida click fraud -menetelmiin nopeasti ja tehokkaasti. Yandex-hakukoneyhtiö on vähentänyt painalluksien merkitystä hakutuloksien järjestämisessä ja antaa sanktion sivulle, jonka havaitaan pyrkivän manipuloimaan hakutuloksia painalluksiin perustuvaa dataa hyödyntäen. (Pevtsov & Volkov, 2013)

Kävijäliikennettä tarjoavat palvelut tarjoavat vierailuja asiakkaan valitsemalle kohdesivustolle muilla sivustoilla vierailuja vastaan. Osa palveluista on ilmaisia ja osa maksullisia. Palvelut voidaan jakaa tavallisiin (generic) palveluihin ja sosiaalisen median promootioon. Toisin kuin sosiaalisen median promootio, tavalliset palvelut mahdollistavat kävijäliikenteen kohdistamisen mille tahansa sivustolle. Tavalliset palvelut voidaan edelleen jakaa kolmeen kategoriaan: manuaalinen, tavallinen automatisoitu (basic autosurf) ja kehittyneempi automatisoitu (advanced autosurf). (Javed et al., 2015)

Manuaaliset kävijäliikennevaihdot vaativat osallistujiltaan *CAPTCHA*-tunnistuksen jokaisen sivulla vierailun yhteydessä. CAPTCHA on testi, jolla pyritään todentamaan, ettei sivulla vierailija ole botti. Tavalliseen automatisoituun kävijäliikennevaihtoon osallistujat saavat linkin, jonka avaaminen selaimessa mahdollistaa automaattisen sivuilla vierailemisen. Sivun JavaScript noutaa automaattisesti vierailtavan sivun tietyn ajan kuluessa ja avaa sivun iframe-tunnisteella. Kehittyneemmässä automatisoidussa kävijäliikennevaihdossa osallistujat saavat omalle laitteellensa ladattavan työkalun. Tavallisen automatisoidun kävijäliikennevaihdon toiminnallisuuden lisäksi työkalu mahdollistaa referrer ja user-agent -kenttien muokkaamisen HTTP-pyynnön yhteydessä. (Javed et al., 2015)

Reputation manipulation service: SEOClerks

Verkkosivujen hakutulossijoitusta voi pyrkiä parantamaan hyödyntämällä verkkosivun mainetta muokkaavia palveluita (reputation manipulation service). Tällaiset palvelut tarjoavat parempaa näkyvyyttä hakukoneiden tulossivujen lisäksi myös sosiaalisessa mediassa, kuten Facebook:ssa tai Twitter:ssä. Palvelut tarjoavat väärennettyjä arvosteluja ja tykkäyksiä ja niiden avulla voi keinotekoisesti tehostaa kohdesivun PageRank-arvoa. Palveluita on tarjolla esimerkiksi Tor-verkossa. Yksi tällaisia palveluja tarjoava sivusto on SEOClerks. Farooqin ja muiden (2017) tutkimuksen teon aikaan sivustolla oli 39 520 palvelua, joista suurin osa oli vilpillisiä (fraudulent). Sivusto myy esimerkiksi linkkejä toisilta sivustoilta, click fraud -palveluja, epäaitoja Instagram-seuraajia, uudelleentviittauksia Twitter-palvelussa ja tykkäyksiä Facebook-palvelussa. Mustahattuiset hakukoneoptimointimenetelmät ovat kuitenkin SEOClerks-sivuston tuottavin palvelu. (Farooqi et al., 2017)

SEOClerks-sivuston tarjoamien palvelujen hinnat olivat Farooqin ja muiden (2017) tutkimuksen aikaan 1-999 USD. Suurin osa palveluista maksoi kuitenkin alle 20 USD. Myyntimäärältään suosituimmat palvelut olivat tykkäyksien ja seuraajien lisäykset sosiaalisen median palveluissa. Nämä palvelut ovat hinnoiltaan pieniä. Esimerkiksi kuudensadan seuraajan lisääminen Instagram-palvelussa maksoi 2 USD. Eniten rahallista tuottoa tuottaneet palvelut olivat hakutulossijoittumista Googlen hakutulossivulla kasvattavat palvelut. Eniten tuottava palvelu oli “Backlinks to improve Google search ranking”, joka

maksoi 29 USD. Palvelua myytiin tutkimuksen aikana 1 364 kertaa, mikä tuotti 39 556 USD. Toinen vastaava palvelu oli "Google X Factor Link Circle For Higher Ranking And Quality Links", jonka hinta oli 57 USD, ja sitä myytiin 550 kertaa tuottaen 31 350 USD. (Farooqi et al., 2017)

SEOClerks-sivustolla oli Farooqin ja muiden (2017) tutkimuksen aikaan 262 909 käyttäjää, joiden joukossa on sekä myyjiä että ostajia. Tutkijat identifioivat 8 861 myyjää ja 33 092 ostajaa. Sama käyttäjä voi olla sekä myyjä että ostaja. Käyttäjät luokitellaan eri tasoille riippuen heidän saavutuksistaan. Tasoja on yhdeksään, ja niiden välillä on mahdollista edetä ansioituneisuuden mukaan. Mitä korkeammalla tasolla käyttäjä on, sitä enemmän etuja käyttäjä saa. Käyttäjä voi esimerkiksi saavuttaa oikeuden alentaa myymiensä palvelujen hintoja, mikä lisää palvelujen suosiota. Sekä myyjinä että ostajina toimivat käyttäjät (dual users) usein ostivat palvelun ja myivät saman palvelun suuremmalla hinnalla. (Farooqi et al., 2017)

Vain suhteellisen pieni *avaintekijöiksi* (key stakeholders) kutsuttu joukko muodostavat merkittävimmän osan palvelujen toimintaa. Ilmiön tunnistaminen on tärkeää sivuston tarjoamien palvelujen mahdollisen alasajon kannalta. Avaintekijät ovat sivuston menestyneimpiä palvelujen myyjiä, jotka ovat liittyneet sivustolle aikaisessa vaiheessa ja he ovat sivuston aktiivisimpia käyttäjiä. SEOClerks-sivuston kaikista käyttäjistä avaintekijöitä oli vain 99, ja he tekivät 56 prosenttia palvelun tuotoista. Avaintekijöillä on oikeus palvelujen esille asettamiseen, mikä lisää palvelujen myyntiä. Yksittäinen avaintekijä myi tutkimuksen aikana parhaimmillaan 1092 palvelua. Avaintekijöiden tekemät myynnit vastasivat 44 prosenttia koko sivuston myynnistä, mikä vastaa 590 357 USD tuottoa. (Farooqi et al., 2017)

SEOClerks-kaltaisten sivuston tarjoamien palvelujen alasajo on haastavaa, koska niiden toimintaa ei tunneta tarpeeksi. Palveluiden alasajo voidaan pyrkiä toteuttamaan esimerkiksi kohdistamalla toimenpiteet oikeusteitse verkkotunnuksen ylläpitoon, mutta prosessi on hidas ja tehoton, sillä palvelu voi nimetä sivustonsa uudelleen tai sijoittaa sen toisaalle. Alasajo voidaan mahdollisesti toteuttaa myös palvelujen käyttämien maksupalvelujen kautta kohdistamalla toimenpiteet oikeusteitse palvelun käyttämille tileille. Koska merkittävä osa palvelun tuotoista on vain pienen joukon aikaansaama, alasajotoimenpiteiden kohdistaminen avaintekijöihin voi olla tehokas menetelmä. (Farooqi et al., 2017)

Private blog networks

Private blog network (PBN, tässä tutkielmassa PB-verkko) koostuu sivustoista, joiden tarkoituksena on osoittaa linkkejä kohdesivustolle ja nostaa kohdesivuston hakutulossijoitusta. PB-verkot ovat osa palvelua, jossa asiakas voi ostaa valitsemalleen kohdesivustolle osoittavia linkkejä. Linkkejä myyvät palvelut tarjoavat eri tasoisia ja eri hintaisia

palveluita. Halvimmille palveluille on ominaista, että palveluntarjoaja luo sivustolle, joka ei ole tämän omistuksessa, profiilin ja lisää profiilisivulleen asiakkaan kohdesivustolle osoittavan linkin. Kalliimmissa palveluissa linkkejä voidaan lisätä sivustoille, jotka ovat palveluntarjoajan omistuksessa, ja PB-verkoissa on kyse tällaisista sivustoista. Tyypillisesti PBN-sivusto on blogisivusto, jolle perustetaan asiakkaan ostamaa linkkiä varten uusi sivu. Sivulle lisätään myös muutakin sisältöä, kuten asiakkaan toiveiden mukaisia avainsanoja. Jotta sivu vaikuttaisi luotettavalta ja hakukoneiden mahdollisuus havaita ohjesääntöjen vastainen toiminta heikentyisi, sisällöstä pyritään tekemään laadukasta. Sisältö voi olla alkuperäistä, tai se voi olla muualta kopioitua. Sisällön luomisessa on voitu käyttää esimerkiksi text-spinning-ohjelmistoa. (Van Goethem et al., 2019)

PB-verkkojen ylläpitäjät omistavat useita verkkotunnuksia, joiden sivustot sisältävät blogeja tai muita sisällönhallintajärjestelmiä. PB-verkon ylläpitäjät pyrkivät usein laajentamaan verkostoaan verkkotunnuksilla, joille on jo entuudestaan osoitettu useita linkkejä aiheeseen liittymättömiltä sivustoilta ja näin nostaa verkoston arvoa hakukoneyhtiöiden näkökulmasta. Linkkien osoittaminen voi myös heikentää kohdesivun ja koko linkkiverkoston sijoittumista, jos hakukone havaitsee ohjesääntöjen vastaisen toiminnan. Pienen verkoston tekemän rikkeen havaitseminen on helppoa, jos jokainen verkoston sivusto osoittaa linkkejä kaikille muille verkoston sivustoille. Palvelua voidaankin käyttää myös negatiiviseen hakukoneoptimointiin, jolloin on tarkoitus osoittaa linkkejä kilpailevalle sivustolle ja saada sen sijoittuminen heikentymään. (Van Goethem et al., 2019)

PB-verkkoihin kuuluvia sivustoja voidaan havainnoida seuraamalla jo PB-verkkoon kuuluvien sivustojen linkkejä ja luokittelemalla sivustoja perustuen niiden sisällöllisiin ja linkkiperustaisiin ominaisuuksiin. PB-verkkoon kuuluvissa sivuissa on havaittavissa yhteisiä rakenteita, joiden perusteella niitä voi luokitella. Tällaiset sivut ovat yleensä luotu käyttäen sisällönhallintajärjestelmää, kuten WordPress, jonka avulla sisältö voidaan luoda nopeasti. Tärkeä tarkastelun kohde ovat sivuilla olevien toiselle verkkotunnukselle osoitettavien (potentiaalisesti linkin ostaneen asiakkaan kohdesivulle osoittavien) linkkien ominaisuudet. Tällaista linkeistä otetaan huomioon esimerkiksi sanamäärä linkin ankkuritekstissä, kohdesivuston Alexa-ranking, linkkien määrä sivustolla ja linkkien URL-osoitteen pituus. Sivulta itsestään huomioon otetaan muun muassa HTML-elementtien rakenne ja sivuston eri sivujen keskinäinen samankaltaisuus. Van Goethem ja muut (2019) havaitsivat PBN-sivuilla olevan muihin sivuihin verrattuna olevan enemmän tekstisisältöä. (Van Goethem et al., 2019)

Sisältöperustaisen luokittelun jälkeen tutkitaan sivun linkkien osoittamien kohdesivujen joukosta PB-verkkoon kuuluvien sivujen määrää. Lisäksi tutkitaan muiden kohdesivustoille linkkejä osoittavien sivustojen joukossa olevien PB-verkkoon kuuluvien sivustojen määrää. PB-verkostoilla on oletettavasti suuri määrä eri asiakkaiden kohdesivustoille osoitettuja linkkejä. Tällaiset kohdesivustot ovat harvoin suosittuja, eivätkä näin

ollen sijoitu korkealle Alexa-ranking-listauksessa. Van Goethemin ja muiden (2019) tekemiin havaintoihin perustuen PB-verkostoihin kuuluvilla PBN-sivuilla on PB-verkkoihin kuulumattomiin sivuihin verrattuna enemmän linkkejä, jotka osoittavat ainutlaatuisiin verkkotunnuksiin, eli todennäköisesti asiakkaiden kohdesivustoihin. Keskimääräisesti uusia linkkejä luodaan PBN-sivustolle yksi kahdessa päivässä. (Van Goethem et al., 2019)

PBN-sivut ovat keskittyneitä lyhytaikaisiin verkkotunnuksiin. Tämä voi johtua siitä, että hakukoneiden havaitessa ohjesääntöjen vastaista toimintaa PB-verkkojen ylläpitäjät ovat ottaneet käyttöön uusia verkkotunnuksia. Ylläpitäjät voivat ostaa vanhentuneita verkkotunnuksia, joilla on jo entuudestaan arvoa hakukoneyhtiöiden näkökulmasta niille osoitettujen linkkien ansiosta. Näin PB-verkolle saadaan nopeasti arvoa. (Van Goethem et al., 2019)

Van Goethem ja muut (2019) havaitsivat, että yli 70% PBN-sivustoista hyödyntää WordPress-alustaa, ja että kohdesivustojen aiheet liittyvät liiketoiminnan ja mainonnan harjoittamiseen. Alle 6% kohdesivuista harjoitti haitallista toimintaa, kuten tietojen kalastelua, haittaohjelmien levitystä, huijauksia ja laittomien palvelujen tarjoamista. (Van Goethem et al., 2019)

Muita keinoja PB-verkostojen tunnistamiseen ovat sivustojen palvelimien sijainnin, WHOIS-informaation ja yhteisten asiakkaiden tutkiminen. Yksi tapa PB-verkon sivustojen yhdistämiseen on *shared hosting*, jossa useat sivustot jakavat palvelimen. WHOIS-informaatioon liittyen Van Goethem ja muut (2019) havaitsivat, että yli puolet PBN-sivustoista käyttivät WHOIS yksityissuojaa. He havaitsivat myös, että asiakkaan ostaessa linkkipalvelun, linkkejä osoitetaan kohdesivustolle useilta verkoston sivuilta. Näin ollen useasta samaan PB-verkkoon kuuluvasta sivustosta on osoitettu linkkejä samalle kohdesivustolle. (Van Goethem et al., 2019)

8 Muut mustahattumenetelmät

Tässä luvussa kuvaillaan cloaking-menetelmää ja uudelleenohjausta sekä niiden havainnointia. Alaluvussa 8.1 pyritään määrittelemään cloaking-menetelmä ja alaluvussa 8.2 kuvaillaan cloaking-menetelmän käyttöä muiden mustahattumenetelmien yhteydessä. Alaluvussa 8.3 kuvaillaan cloaking-menetelmän toteutusta ja alaluvussa 8.4 cloaking-menetelmän havainnointia. Alaluvussa 8.5 pyritään määrittelemään uudelleenohjaus ja kuvaillaan, miten uudelleenohjaus voidaan toteuttaa. Alaluvussa 8.6 kuvaillaan uudelleenohjauksen havainnointia. Alaluvussa 8.7 kuvaillaan uudelleenohjaavien sivujen muodostamaa vertaisverkkoa.

8.1 Cloaking-menetelmän määritelmä

Cloaking on piilotusmenetelmä, jota hyödyntäen hakukoneen ryömijälle ja selaimen käyttäjälle esitetään verkkosivulla vieraillessaan eri sisältö. Verkkotunnuksen esittämä sisältö riippuu sivulla vierailijan identiteetistä. (Wang et al., 2014; Fox 2008) Cloaking-menetelmän hyödyntäjän tulee tunnistaa sivun vierailijan (client) identiteetti. Yksi verkkotunnus voi palauttaa eri HTML-dokumentin riippuen vierailijan identiteetistä. Hakukoneen ryömijälle voidaan palauttaa esimerkiksi optimoitu dokumentti, kun taas selainta käyttävälle vierailijalle palautetaan käyttäjäystävällinen sisältö. (Gyongyi & Garcia-Molina, 2005) Toisaalta hakukoneen ryömijälle esitettävä sisältö voi olla neutraalia, kun taas selainta käyttävälle vierailijalle esitettävä sisältö on haitalliseksi koettua (Wang et al., 2011) tai sisältää runsaasti mainoksia (Patil et al., 2021).

8.2 Cloaking muiden menetelmien yhteydessä

Cloaking-menetelmän käytön yhteydessä saatetaan käyttää muita mustahattuisia hakukoneoptimointimenetelmiä. Cloaking-sivut saattavat esimerkiksi olla yhteydessä linkkifarmeihin ja useat cloaking-sivut hyödyntävät uudelleenohjausmenetelmää (jota käsitellään alaluvussa 8.5) ja täyteavainsanoja. Wangin ja muiden (2011) tutkimista cloaking-sivuista osa hyödynsi sivun aiheesta riippumattomia suosittuja avainsanoja saavuttaakseen laajaa kävijäliikennettä. Osa cloaking-sivuista taas hyödynsi aiheeseen liittyviä avainsanoja, jolloin sivujen ohjesääntöjen vastainen toiminta pysyi pidempään hakukoneiden huomaamattomissa. Kun cloaking-sivua on optimoitu tietyillä avainsanoilla ja selaimen käyttäjä vierailee sivulla kyseisillä hakusanoilla suoritetun haun kautta, selainta käyttäneelle vierailijalle saatettiin esittää haitallista sisältöä. (Wang et al., 2011)

Cloaking-ominaisuus voi olla injektoituna hyvälaatuiselle sivulle sivun ylläpitäjän tietämättä. Sivun ylläpitäjän voi olla haastavaa havaita injektiota, sillä selaimen käyttäjä voi saada poikkeavan sisällön vasta, kun sivulle saavutaan hakutuloksien kautta. Cloaking-ominaisuus voi säilyä injektoiduilla sivuilla useita päiviä. Wangin ja muiden (2011) tutkimassa aineistossa suurimmassa osassa injektoituja sivuja cloaking-ominaisuus säilyi yli viikon. Injektiot voivat aiheuttaa vahinkoa sekä sivun vierailijoille että sivulle itselleen. Injektiot voivat esimerkiksi kerryttää kohdesivulle vierailijoita asentaakseen näiden laitteille haittaohjelmia. Haitallisen sisällön levittämisen vastatoimena Googlen Safe Browsing -ominaisuus merkitsee hakutuloksia haitallisiksi havaitessaan niiden yhteydessä esimerkiksi tietojen kalastelua tai haittaohjelmien levitystä. Haitalliseksi merkityn sivun maineen palauttaminen voi olla työlästä. (Wang et al., 2011)

8.3 Cloaking-menetelmän toteutus

Vierailijan identiteetin tunnistamiseksi cloaking-menetelmän harjoittajalla voi olla informaatio hakukoneiden käyttämistä IP-osoitteista, jota hän voi käyttää pyynnön esittäjän identifioimiseen. Cloaking-sivun palvelin voi tunnistaa pyynnön tekevän sovelluksen perustuen HTTP request message -viestissä olevaan user-agent -kenttään. (Gyongyi & Garcia-Molina, 2005) Seuraa esimerkki, jossa HTTP request message -viestissä on Microsoft Internet Explorer 6 –selaimen käyttämä user-agent -nimi:

> "GET /db pages/members.html HTTP/1.0 Host: www-db.stanford.edu User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
> (Gyongyi & Garcia-Molina, 2005)

User-agent -nimet eivät ole tarkoin standardoituja ja pyynnön tekevä sovellus voi itse määrittää request message -viestin sisällön. Tästä huolimatta hakukoneiden ryömijät identifioivat itsensä, sillä ryömijöille voidaan palauttaa eri sisältöä kuin selaimelle myös legitiimissä tarkoituksessa. Sivulta voidaan poistaa ryömijälle tarkoitetusta sisällöstä esimerkiksi mainokset, navigointilinkit sekä visuaalinen sisältö, mikä nopeuttaa sivun prosessointia ja indeksointia. (Gyongyi & Garcia-Molina, 2005)

Taulukossa 4 kuvaillaan viisi menetelmää sivun vierailijan identiteetin tunnistamiseen. Kun vierailijan identiteetti on tunnistettu, voidaan verkkosivusta esittää haluttu versio.

Menetelmä	Kuvaus
IP-cloaking	Vierailijan identiteetin (ryömijä tai selain) voi tunnistaa pyynnön (request) lähettäjän IP-osoitteesta, jolloin hakukoneiden identiteetti on havaittavissa.
User-agent-cloaking	Hakukoneilla on tunnistettava user-agent-string (esimerkiksi: Google-bots' UA: Googlebot/2.1 (+http://www.googlebot.com/bot.html)). HTTP-pyynnön otsikon user-agent -kentästä on mahdollista erotella pyynnön lähettäjän identiteetti.
Referrer cloaking	HTTP-pyynnön otsikon referrer-kenttää tutkimalla voi erottaa, minkä URL:n kautta vierailija saapuu sivustolle.
Repeat-cloaking	Sivusto määrittää, onko vierailija vieraillut sivustolla aiemmin. Tämä informaatio perustuu esimerkiksi evästeisiin.
JavaScript redirection cloaking	Uudelleenohjauksen toteuttava skripti on sisällytetty sivustoon. Ryömijät eivät lue skriptejä kuten selaimet.

Taulukko 4: Viisi menetelmää sivun vierailijan identiteetin tunnistamiseen (Deng et al., 2013)

8.4 Cloaking-menetelmän havainnointi

Cloaking-menetelmää hyödyntäviä sivuja voi pyrkiä havainnoimaan vertailemalla sivun ryömijälle ja selaimelle esitettäviä versioita. Koska dynaamisesti generoitu tai usein päivitetty sisältö voi tulla virheellisesti luokiteltua mustahattuiseksi, vertailuun on syytä käyttää useampaa kuin kahta versiota. Sivun versioiden vertailu perustuu termien, linkkien ja tunnisteiden poikkeavuuteen. Vertailun kohteena voi olla esimerkiksi toisistaan poikkeavat termit ja linkit, termien ja linkkien lukumäärien poikkeamat sekä termien esiintymistiheyksien poikkeamat. Tunnisteet ovat hyödyllinen vertailun kohde, sillä ne eivät muutu yhtä usein, kuin linkit ja termit. (Ghiam, 2012)

Vertailu voidaan suorittaa esimerkiksi kolmen version välillä, joista kaksi ovat ryömijälle esitettäviä ja yksi selaimelle esitettävä versio. Jos selaimelle ja toiselle ryömijälle esitettävän version välinen ero on suurempi kuin kahden ryömijälle esitettävän version välinen ero, kyseessä voi olla cloaking-menetelmää hyödyntävä sivu. Vertailua voidaan edelleen kehittää hyödyntämällä kahta selaimelle esitettävää versiota (B1 ja B2) ja kahta ryömijälle esitettävää versiota (C1 ja C2). Näin ollen esimerkiksi termejä voidaan vertailla laskemalla B1 ja B2 yhteiset termit, joita ei esiinny C1 ja C2 -versioissa sekä C1 ja C2 yhteiset termit, joita ei esiinny B1 ja B2 -versioissa. Nämä lukemat yhteen laskettuna ja niiden ylittäessä tietyn raja-arvon sivu voidaan tulkita mustahattuiseksi. (Ghiam, 2012)

Dengin ja muiden (2013) menetelmä cloaking-sivujen havainnoimiseksi koostuu kahdesta komponentista, joista ensimmäinen ryömii verkkoa (data crawling component) ja toinen tutkii sivujen samankaltaisuutta (similarity detecting component). Ensimmäinen komponentti suorittaa kolme operaatiota. Se etsii suositun hakusanan, jonka jälkeen se suorittaa verkkohaun ja kerää tuloksina saadut URL-osoitteet. Seuraavaksi se tutkii näiden sivustojen HTML-tiedostot sekä hakukoneen ryömijälle että selaimelle esitettävinä versioina. Tästä saatu informaatio siirtyy samankaltaisuutta tutkivan komponentin käsiteltäväksi, joka niin ikään suorittaa kolme operaatiota. Komponentti tutkii sivun sisältöä perustuen tekstiin, tunnisteisiin ja URL-osoitteeseen. Jos sivun kahden eri version (ryömijän ja selaimen) laskettu samankaltaisuus alittaa tietyt raja-arvot, voidaan olettaa, että sivulla on hyödynnetty cloaking-menetelmää. (Deng et al., 2013)

Deng ja muut (2013) kykenivät menetelmällään havaitsemaan taulukossa 4 kuvattuja menetelmiä hyödyntäviä sivuja, lukuun ottamatta repeat cloaking -menetelmää hyödyntäviä sivuja, sillä tutkittaville sivustoille ei tehty URL-pyyntöä useampaa kertaa samana vierailijana. Deng ja muut (2013) havaitsivat, että useimmiten cloaking-sivustot hyödyntävät samanaikaisesti useampaa cloaking-menetelmää. (Deng et al., 2013)

Wang ja muut (2011) havainnoivat cloaking-sivuja vierailemalla tutkittavilla kohdesivuilla kolme kertaa ja vertailemalla näin samasta sivusta saatuja versioita. Sivulla vierailtiin selaimella ja ryömijällä sekä uudelleen selaimella, joka ensimmäisestä kerrasta poiketen ei ole saapunut sivulle painamalla hakutulossivun hakutulosta. Vierailijan identiteettiä muutettiin muokkaamalla HTTP-pyynnön otsikon user-agent- ja referrer-kenttiä. Wang ja muut (2011) havaitsivat, että 35% heidän tutkimistaan cloaking-sivuista käyttivät user-agent cloaking -menetelmää. Tällaiset sivut olivat harvoin haitallisia käyttäjälle, mutta sivut, jotka hyödynsivät sekä user-agent ja referrer-cloaking -menetelmiä, olivat lähes aina haitallisia. (Wang et al., 2011)

Hakutulossivulla esitetään usein katkelma (snippet) sivun sisällöstä hakutuloksen yhteydessä. Katkelma osoittaa tuloksen relevanssin käyttäjälle esittämällä sivusta kyselyyn viittaavan kohdan. Katkelmaa voidaan hyödyntää cloaking-sivujen havainnoimiseksi ver-

tailemalla katkelman ja sivun varsinaisen tekstien vastaavuutta, sillä tekstien ero voi viitata cloaking-menetelmän hyödyntämiseen. On kuitenkin huomioitava, että poikkeavuus voi johtua myös sivulle tehdystä päivityksestä. Vertailun kohteena onkin syytä käyttää myös sivujen HTML-rakennetta, sillä vaikka sivun tekstisisältöä päivitettäisiinkin, sivun rakenne muuttuu harvoin. Jos siis sivun rakenteessa havaitaan eroa, se voi viitata cloaking-menetelmän hyödyntämiseen. (Wang et al., 2011)

8.5 Uudelleenohjausmenetelmän määritelmä ja toteutus

Uudelleenohjaus (redirection) on menetelmä, jossa verkkosivulla vierailun yhteydessä selaimen käyttäjä ohjataan automaattisesti toiselle sivulle (Li et al., 2014). Menetelmää voidaan hyödyntää muiden mustahattumenetelmien piilottamiseen (Gyongyi & Garcia-Molina, 2005; Somani & Suman, 2011). Uudelleenohjaava sivu voidaan hyödyntää mustahattuhakukoneoptimointiin, eikä selaimen käyttäjä näe kyseistä sivua, sillä uudelleenohjaus tapahtuu nopeasti usein selaimen käyttäjän huomaamatta. (Gyongyi & Garcia-Molina, 2005). Kävijäliikennettä voidaan ohjata kohdesivustolle useista osoitteista (site mirroring) (Patil et al., 2021)

Uudelleenohjaus voidaan toteuttaa hyödyntämällä HTML-dokumentin header-osion refresh-metatunnistetta. Refresh time -arvo voidaan asettaa nollaksi ja refresh URL-osoitteeksi voidaan asettaa kohdesivu, jolloin uudelleenohjaus tapahtuu heti vierailijan saapuessa sivulle:

```
<meta http-equiv="refresh" content= "0;url=target.html">
```

(Gyongyi & Garcia-Molina, 2005)

Yllä mainittu meta refresh -menetelmä on kuitenkin helposti havaittavissa, sillä ryömijät huomioivat metatunnisteen sisällön (Gyongyi & Garcia-Molina, 2005). Useat hakukoneyhtiöt ovat poistaneet indeksistään sivut, jotka hyödyntävät meta refresh -ominaisuutta. Tästä johtuen uudelleenohjaava ominaisuus sisällytetään usein sivun dynaamiseen sisältöön, kuten JavaScriptiin. (Malaga, 2008) Ryömijä huomioi skriptin metatiedot, kuten header-tunnisteen ja metatunnisteen, mutta ei kohdetta, johon skripti ohjaa sivun vierailijan (Somani & Suman, 2011). Näin ollen hakukoneet eivät välttämättä havaitse uudelleenohjausominaisuutta, eivätkä tulkitse sivua haitalliseksi (Li et al., 2014). Seuraa esimerkki, jossa uudelleenohjausominaisuus on asetettu skriptiin, jota ryömijät eivät huomioi:

```
<script language="javascript"> <!--location.replace("target.html")
--></script>
```

(Gyongyi & Garcia-Molina, 2005)

Uudelleenohjauksen yhteydessä voidaan hyödyntää useita lyhytikäisiä verkkotunnuksia (throwaway domains), joita hyödynnetään vain kävijäliikenteen kasvattamiseen (Somani & Suman, 2011). Kohdesivulle ohjaavia doorway-sivuja voi olla satoja, jotka ovat optimoituja eri avainsanoja hyödyntäen (Malaga, 2008). Uudelleenohjaavat sivut optimoidaan esimerkiksi suosituilla avainsanoilla (Somani & Suman, 2011). Tässä tapauksessa doorway-sivuilla on vain vähän sisällöllistä merkitystä, mutta ne sisältävät runsaasti avainsanoja, jotka voidaan sijoittaa näkymättömiin selaimen käyttäjältä (Patil et al., 2021). Uudelleenohjausta voidaan hyödyntää myös cloaking-menetelmän yhteydessä. (Somani & Suman, 2011; Li et al., 2014) Cloaking-menetelmää hyödyntäen hakukoneet eivät välttämättä havaitse uudelleenohjausta (Somani & Suman, 2011).

Uudelleenohjaus on suosittu menetelmä haitallisen verkkosisällön levittämisessä johtuen sen huomaamattomuudesta. Uudelleenohjaava ominaisuus voidaan injektoida haavoittuvalle sivulle. Haavoittunut sivu ei välttämättä ole uudelleenohjaavaa ominaisuutta lukuun ottamatta haitallinen, sillä uudelleenohjaava ominaisuus voidaan injektoida haavoittuvalle sivulle sivun ylläpitäjän huomaamatta. Uudelleenohjaus voidaan toteuttaa injektoimalla HTML-tiedostoon uudelleenohjausominaisuuden sisältävä tunniste. Injektio voidaan asettaa esimerkiksi frame-, iframe- tai script-tunnisteisiin, jolloin uudelleenohjaus on käyttäjälle huomaamatonta. Uudelleenohjaus voidaan toteuttaa myös injektoimalla skripti JavaScript-tiedostoon (JS-tiedosto) tai HTML-tiedostoon. Skriptit ohjaavat vierailijan selaimen haluttuun sijaintiin, tai luovat dynaamisesti HTML-tunnisteen, joka aiheuttaa haitallisen sisällön lataamisen selaimella. Koska hakukoneet eivät indeksoi JavaScript-tiedostoja, niiden injektointi on suosittua. Injektoinnin yhteydessä voidaan hyödyntää menetelmiä, joilla pyritään piilottamaan uudelleenohjausominaisuus. Piilotusmenetelmien käyttö itsessään voi kuitenkin edistää uudelleenohjauksen havainnointia ja näin ollen esimerkiksi iframe-tunnistetta ei aina pyritä piilottamaan uudelleenohjauksen yhteydessä. (Li et al., 2014)

8.6 Uudelleenohjausinjektioiden havainnointi

Injektioita toteuttavissa uudelleenohjauskampanjoissa on havaittavissa tunnistettavia piirteitä, joita voidaan hyödyntää uudelleenohjauksen havainnoimiseksi. Uudelleenohjauskampanjoissa injektioita toteutetaan runsaasti eri sivustoille lyhyessä ajassa. Injektioita toteutetaan “sokeasti” eri kohdesivustojen JavaScript- ja HTML-tiedostoihin, eikä injektioita räätälöidä erikseen hyökättäville sivustoille. Näin kampanjaa pyritään tehostamaan kasvattaen kampanjan vaikutusta. Samankaltaista skriptiä injektoidaan useisiin tiedostoihin ja skriptit saattavat olla jopa täysin samankaltaisia eri kohteisiin injektoituna. Uudelleenohjausskriptien samankaltaisuutta voidaan näin ollen hyödyntää niitä havainnoitaessa. (Li et al., 2014)

Uudelleenohjausinjektioiden yhteydessä tehdyt muutokset JavaScript-tiedostoihin ovat pieniä, millä pyritään pitämään injektio mahdollisimman huomaamattomana. JavaScript-tiedoston toimintaa ei pyritä muuttamaan muilta osin. Hyökkäyksen kohteena olevat JavaScript-tiedostot ovat usein täysin tai lähes alkuperäisessä muodossaan olevia JavaScript-kirjastoja (JS-lib-tiedostoja), joiden alkuperäiset versiot ovat julkisesti saatavilla. Näin ollen epäiltyä injektion sisältävää JavaScript-tiedostoa voidaan vertailla alkuperäiseen JS-lib-tiedostoon. JS-lib-tiedostoja voi olla muokattu myös ei-haitallisessa tarkoituksessa, mutta niissä ei usein ole uudelleenohjausominaisuutta. Uudelleenohjausominaisuuden löytyessä kyseessä voi olla haitallisessa tarkoituksessa injektoitu tiedosto. Kun haavoittuneesta tiedostosta on löydetty uudelleenohjaava skripti, vastaavanlaista skriptiä voidaan etsiä laajemmasta joukosta tiedostoja. (Li et al., 2014)

8.7 Uudelleenohjaavien sivujen muodostama vertaisverkko

Uudelleenohjauskampanjoissa haavoittuneet sivut voivat muodostaa *vertaisverkon* (peer-to-peer, P2P), jonka tarkoituksena on suojata haavoittuneita sivustoja tulemasta havainnoiduiksi. Uudelleenohjausinfrastruktuurin rakenteessa on tavallisesti kolme osaa: sisäänpääsynä toimiva haavoittunut sivu, haitallinen uudelleenohjausjärjestelmä (esimerkiksi Traffic Direction Systems) ja kohde (esimerkiksi drive by download -sivut). Haavoittuneet sivut muodostavat verkoston, jota pitkin vierailija kulkee sivulta toiselle ennen kuin päätyy kohteeseen, joka voi myös olla haavoittunut sivu, josta haitallinen sisältö toimitetaan vierailijan selaimeen. Tällainen verkko vaikeuttaa uudelleenohjauksien seuraamista. (Li et al., 2014)

Haavoittuneet sivut näyttelevät jotakin kolmesta roolista: *uudelleenohjaava solmu* (relay node), *poistumissolmu* (exit node) sekä *kohdesolmu* (target node). Uudelleenohjaavien solmujen tehtävänä oli lähettää vierailija joko toiselle uudelleenohjaavalle solmulle tai poistumissolmulle. Poistumissolmut, jotka toimivat myös uudelleenohjaajina, voivat ohjata vierailijan ulos verkosta luomalla dynaamisesti kohdesolmun ja ohjaamalla vierailijan kyseiseen kohteeseen. Kohdesolmut ovat myös haavoittuneita sivuja, joita voidaan hyödyntää esimerkiksi haittaohjelmien levittämiseen vierailijoille. Poistumissolmut ovat pysyvämpiä verrattuna uudelleenohjaaviin solmuihin, joiden verkkosivut vaihtelevat usein. Lisäksi poistumissolmut vaihtavat usein osoitteita, joille ne ohjaavat vierailijat. Tämä viittaa strategiaan, jossa hyödynnetään suurta määrää uhrattavissa olevia uudelleenohjaavia solmuja, jotka osoittavat liikenteen suhteellisen pysyville ja hyvin suojelluille poistumissolmuille. Poistumissolmut taas valitsevat dynaamisesti hyväksikäytettäviä palvelimia, joille vierailijat ohjataan. (Li et al., 2014)

Tutkiessaan uudelleenohjaavia vertaisverkkoja Li ja muut (2014) havaitsivat verkon hylkivän tutkijoiden käyttämää ryömijää. Lin ja muiden (2014) tutkima vertaisverkko oli toiminnassa vähintään viisi ja puoli kuukautta. Haavoittunut sivu sisälsi infektion keskimäärin viisitoista päivää, mediaanin ollessa kahdeksan päivää, ja maksimiajan ollessa 162

päivää. Li ja muut (2014) havaitsivat vertaisverkkoon kuuluvien haavoittuneiden sivujen pysyvän injektoituna pidempään kuin vertaisverkkoon kuulumattomat injektoidut sivut, mikä viittaa vertaisverkon heikentävän havainnoiduksi tulemista. Li ja muut (2014) havaitsivat useiden haavoittuneiden sivujen olleen luotuja alustoilla, kuten WordPress, Joomla ja Plesk. (Li et al., 2014)

9 Mustahattuhakukoneoptimoinnin vastatoimet

Tässä luvussa kuvaillaan mustahattuhakukoneoptimoinnin vastatoimia. Alaluvussa 9.1 kuvaillaan mustahattuhakukoneoptimoinnin haitallisuutta perusteena mustahattuhakukoneoptimoinnin vastatoimille. Lisäksi alaluvussa kuvaillaan mustahattumenetelmien havainnoinnin haasteita. Alaluvussa 9.2 kuvaillaan, miten mustahattumenetelmiä hyödyntävien verkkosivujen ominaisuuksia voidaan hyödyntää mustahattuhakukoneoptimoinnin havainnoimiseen. Alaluvussa 9.3 kuvaillaan itseoppivia algoritmeja vastatoimien välineenä ja alaluvussa 9.4 kuvaillaan TrustRank- ja BadRank-arvojen hyödyntämistä vastatoimien välineenä. Alaluvussa 9.5 kuvaillaan käyttäjäkokemuksesta saadun informaation hyödyntämistä vastatoimien välineenä ja alaluvussa 9.6 kuvaillaan Googlen algoritmeja vastatoimien välineenä. Alaluvussa 9.7 kuvaillaan peliteoriaa negatiivisen hakukoneoptimoinnin vastatoimena.

9.1 Mustahattuhakukoneoptimoinnin haitallisuus ja havainnoinnin haasteet

Mustahattuhakukoneoptimointi voi olla haitallista sekä hakukoneille että hakukoneiden käyttäjille. Mustahattusivut tuhlaavat tiedonhakijan aikaa, sillä ne eivät tarjoa relevanttia informaatiota. Tämä taas voi aiheuttaa epäluottamusta hakukoneita kohtaan. Mustahattusivuilla vieraileminen voi aiheuttaa esimerkiksi haittaohjelman leviämisen käyttäjän laitteelle. Mustahattusivut hukkaavat hakukoneiden resursseja, sillä verkon ryömiminen kuluttaa verkon kaistanleveyttä, prosessoiminen prosessorin syklejä ja indeksointi tallennustilaa. (Ghiam, 2012)

Hakukoneiden ryömijöitä harhauttavat menetelmät ja luotettujen sivujen osoittamat linkit mustahattusivuille lisäävät mustahattusivujen havainnoinnin haasteellisuutta (Somani & Suman, 2011). Mustahattusivujen sisältö voi olla kopioitua laadukkailta sivustoilta, tai se voi olla sivun muiden ominaisuuksien ohella piilotettu cloaking-menetelmää hyödyntäen. Mustahattusivujen muodostamien linkkien verkosto voi olla monimutkainen (Svore et al., 2007). Mustahattusivujen manuaalisen havainnoinnin tehottomuus luo tarpeen nopeammalle ja tarkemmalle havainnoinnille (Ghiam, 2012). Toisaalta, havainnoinnin haasteellisuus luo tarpeen havainnoinnin tehostamiselle ryömijöiden suorittaman havainnoinnin lisäksi (Somani & Suman, 2011). Sivustojen sijoituksen alenemisen tai hakukoneindeksistä kokonaan poistamisen välttämiseksi on syytä välttää tavallisten sivujen virheellistä tulkitsemista mustahattuiseksi (Svore et al., 2007).

Relevanttien hakutuloksien järjestelemisen kannalta on tärkeää, että mustahattumenetelmien käyttö tunnistetaan. Pelkästään hakutuloksia järjestelevä järjestelyalgoritmi saattaa olla kykenemätön havaitsemaan mustahattumenetelmiä ja sijoittaa niitä hyödyntävät sivut korkealle hakutulossijoille. Sen sijaan mustahattumenetelmien havainnointiin tarkoitettu algoritmi havaitsee mustahattumenetelmät perustuen ominaisuuksiin, jotka tekevät mustahattusivusta poikkeavan muihin sivuihin nähden. Esimerkiksi suuri avainsanamäärä saattaa tehdä sivustosta relevantin tavalliselle järjestelyalgoritmille, mutta mustahattumenetelmien tunnistamiseen keskittyvä algoritmi havaitsee suuren avainsanamäärän aiheuttavan poikkeaman hyvänlaatuisiin sivustoihin nähden. Mustahattusivujen havainnoimisen edistämiseksi on hyödynnettävä järjestelyalgoritmin lisäksi mustahattumenetelmien tunnistamiseen tarkoitettua algoritmia. (Svore et al., 2007)

9.2 Mustahattusivujen ominaisuuksien hyödyntäminen havainnoimisessa

Mustahattusivujen havainnoimiseksi tutkitaan niitä sivun ominaisuuksia, joihin mustahattumenetelmät usein perustuvat. Ominaisuuksia tutkimalla sivu voidaan luokitella mustahattuiseksi. Tutkimuksen kohteena voi olla esimerkiksi sivulla käytettyjen termien lukumäärä ja tiheys sekä koko sivuston sivujen keskimääräinen sanamäärä. Lisäksi voidaan tutkia sivulla vierailujen määrää sekä sivulle tehtyjen muokkauksien ja sivulla ryömimisen ajankohtia. Sivua voidaan myös tutkia suhteessa esitettyyn kyselyyn vertaamalla kyselyssä esiintyviä termejä sivun avainsanoihin. (Svore et al., 2007)

Useat mustahattusivut ovat luotuja samoja alustoja hyödyntäen, joten HTML-sivun rakennetta voidaan hyödyntää mustahattusivun tunnistamiseen. Lisäksi tunnistamiseksi voidaan tutkia HTML-sivun tunnisteiden tyyppejä ja lukumäärää. Mustahattusivun tunnistamiseksi tulisi tutkia myös sivun dynaamisia osia, kuten JavaScriptiä. (Liu et al., 2020)

Tutkittavalta sivulta voidaan selvittää sen sisältämien linkkien lukumäärä. Mustahattusivuille on ominaista, että sivulle muualta osoitettujen linkkien ankkuriteksti ei havainnollista kohdesivun sisältöä. Tästä syystä sivulle osoittavan linkin ankkuritekstin ja kohdesivun tekstin samankaltaisuutta voidaan hyödyntää mustahattusivun tunnistamiseen. (Liu et al., 2020)

Mustahattusivuja voidaan havainnoida tutkimalla mustahattusivujen linkkien välityksellä muodostunutta verkkoa. Mustahattusivulle linkkejä osoittavat sivut voidaan selvittää hyödyntämällä hakukoneen links:target -ominaisuutta. Linkkejä osoittavat sivut tutkitaan ja mustahattusivun löytyessä sama prosessi voidaan toistaa edelleen, jolloin käsitys mustahattusivustojen linkkien välisestä verkosta täydentyy. (Somani & Suman, 2011)

9.3 Itseoppivat algoritmit

Mustahattuhakukoneoptimoinnin vastaisen toiminnan edellytyksenä on mustahattumenetelmien tunnistaminen ja niitä hyödyntävien sivujen havainnoiminen. Mustahattumenetelmien tunnistaminen perustuu sivun sisällöllisten ja linkkiperustaisten ominaisuuksien tutkimiseen sekä mustahattumenetelmille tyypillisten säännönmukaisuuksien havaitsemiseen. Mustahattumenetelmien kehittyminen ja niitä hyödyntävien sivujen runsaus luovat tarpeen automaattiselle mustahattusivujen havainnoimiselle, sillä manuaalinen havainnointi on hidasta ja tehotonta. Koneoppimisen hyödyntäminen tehostaa havainnointia, ja kirjallisuudessa tutkitaan itseoppivia algoritmeja. Algoritmit luokittelevat verkkosivuja perustuen niiden ominaisuuksiin ja kykenevät itse luomaan säännönmukaisuuksia mustahattusivujen tunnistamiseksi ja parantamaan luokittelukykyään. (Li, 2014; Kumar et al., 2016) Algoritmit havainnoivat mustahattusivuja ryömimällä niiden muodostamia linkkiverkostoja (Somani & Suman, 2011; Taweesiriwate et al., 2012) Esimerkiksi Taweesiriwaten ja muiden (2012) algoritmi ryömii linkkiverkostoa useita kertoja ja hyödyntää toistuvuutta luokittelunsa tarkkuuden edistämiseksi. Algoritmien luokittelukykyä on pyritty edistämään myös lisäämällä niiden tunnistamien mustahattumenetelmien määrää. Esimerkiksi Kumarin ja muiden (2016) algoritmi luokittelee sivut sisältö- ja linkkiperustaista mustahattumenetelmää hyödyntäviin sekä cloaking-menetelmää tai useita menetelmiä hyödyntäviin sivuihin (Kumar et al., 2016).

9.4 *TrustRank* ja *BadRank*

TrustRank- ja *BadRank* -arvoilla pyritään vähentämään linkkiperustaista mustahattuhakukoneoptimointia (Kumar et al., 2016). BadRank-arvo on PageRank-arvon kaltainen, mutta käänteinen toimintalogiikaltaan. Sivun BadRank-arvo kasvaa sen osoittaessa linkkejä sivuille, joilla on korkea BadRank-arvo. Kaupalliset hakukoneyhtiöt ovat mahdollisesti hyödyntäneet BadRank-menetelmää linkkifarmien vaikutuksien vähentämiseen. (Wu & Davison, 2005) Myös TrustRank toimii samankaltaisesti kuin PageRank. TrustRank-menetelmän oletuksena on, että hyvälaatuiseksi todetut sivustot harvoin osoittavat linkkejä mustahattuiseille sivustoille. Näin ollen sivun arvo nousee, jos sille on osoitettu linkki sivulta, jolla on korkea TrustRank-arvo. (Wu & Davison, 2005; Ghiam, 2012) TrustRank ei kuitenkaan huomioi hyvälaatuisten sivujen mustahattusivuille osoittamia linkkejä (Ghiam, 2012).

9.5 Käyttäjäkokemus järjestelyn perusteena

Verkkosivun sisältämien mainoksien runsaus voi vaikuttaa negatiivisesti sivun käyttäjäkokemukseen. Kun mainoksien määrä on haitallinen, sitä voidaan kutsua *aggressiiviseksi mainostamiseksi*. Käyttäjäkokemusta voidaan mitata mittaamalla sivulla vietettyä aikaa (dwell time). Yandex-hakukone on hyödyntänyt aggressiivista mainontaa ja sivulla vie-

tettyä aikaa havainnoivia algoritmeja, joiden perusteella samaa sisältöä tarjoavat kilpailevat sivustot voidaan järjestää niiden käyttäjäkokemuksen ja niiden sisältämän mainonnan perusteella. Tämä vähentää esimerkiksi doorway-sivujen kannattavuutta, jotka ovat optimoituja kopioidulla sisällöllä ja joissa käytetään aggressiivista mainontaa. (Pevtsov & Volkov, 2013)

9.6 Google Zoo

Google Zoo sisältää Googlen algoritmeja, joita hyödyntäen Google pyrkii pitämään hakutuloksensa laadukkaina ja relevantteina sekä heikentämään huonolaatuisten sivujen sijoittumista. Google Zoo sisältämät algoritmit ovat Panda, Penguin, Hummingbird, Fred, RankBrain, Bert, Pigeon ja Possum. (Patil et al., 2021)

Panda- ja Penguin-algoritmien tarkoituksena on saada hakutuloksista relevantimpia ja rangaista sivustoja, jotka hyödyntävät mustahattumenetelmiä. Panda-algoritmi alentaa sellaisten sivujen sijoittumista, joiden sisältö on huonolaatuista tai kopioitua ja sisältävät huomattavasti mainontaa. Panda-algoritmin tarkoituksena on arvostaa sisällön laatua (informatiivisuutta), ei määrää. (Jha & Saraswat, 2018) Sivun sisältö on puutteellista, jos siitä on vain vähän tai ei ollenkaan hyötyä vierailijalle. Puutteellisen sisällön sivuja voivat olla esimerkiksi doorway-sivut, huonolaatuiset oheissivut (partner pages), tai vain vähän sisältöä sisältävät sivut. Rangaistuksen voi saada myös heikosta käyttäjäkokemuksesta. Jos sivua ei ole suunniteltu käyttäjäystävälliseksi, sivulla navigointi voi olla haastavaa, eikä käyttäjä kykene selaamaan sivuston sisältöä nopeasti. Myös verkkohaussa esitetylle kyselylle epärelevantti sisältö voi aiheuttaa heikon käyttäjäkokemuksen. Myös liiallisesta avainsanojen tai niiden synonyymien käytöstä voidaan rangaista. Vuodesta 2017 lähtien toiminnassa ollut Fred-algoritmi lisää Pandan toiminnallisuutta. Sen tehtävänä on havainnoida esimerkiksi sisällön huonolaatuisuutta, mainoskeskeisyyttä, asiakeskeisyyden vähäisyyttä ja epärelevanssia. (Patil et al., 2021)

Penguin-algoritmin tarkoituksena on kannustaa sivustovastaavia poistamaan sivuiltaan huonolaatuiset linkit (Jha & Saraswat, 2018). Lisäksi Penguin-algoritmin avulla Google pyrkii ehkäisemään liiallista avainsanojen käyttöä, sisällön piilottamista ja arvokkaiden sivustojen sisällön kopiointia (Patil et al., 2021).

Patilin ja muiden (2021) mukaan sivun sisällöstä 28 prosenttia ollessa kopioitua sisältö on Googlen ohjesääntöjen vastaista. Sisältö on ohjesääntöjen vastainen myös, kun puolet sisällöstä ei liity sivuston pääasiaan. Lisäksi kolmesanaiset lauseet tulkitaan ylioptimoinniksi. (Patil et al., 2021)

Hummingbird-algoritmi käsittelee keskustelevia (conversational) kyselyjä. Algoritmi antaa sivuston kehittäjälle mahdollisuuden hyödyntää sivun hakukoneoptimoinnissa luonnollista kieltä pelkkien avainsanojen lisäksi. Hummingbird huomioi myös synonyymit, mikä tuottaa hakutuloksiin enemmän hakijan esittämän kyselyn teemaan liittyviä tuloksia. Hummingbird-algoritmin osana on myös koneoppimista hyödyntävä RankBrain-

algoritmi. RankBrain-algoritmin tarkoituksena on kehittää hakukoneen “ymmärrystä” verkkohakukyselyjä kohtaan. (Patil et al., 2021)

Panda, Penguin ja Hummingbird -algoritmit ovat olleet käytössä jo 2010-luvun alusta ja ne luovat perustan nykyisin käytössä oleville Googlen algoritmeille. Niiden pohjalta on luotu uusia algoritmeja, jotka jalostavat vanhojen algoritmien toimintaa. Bert-algoritmi on Panda-, Hummingbird- ja RankBrain -algoritmien kulminaatio. Se havainnoi sisällön puutteita, kuten kirjoitusasun, asiasisällön sekä yhteneväisen sisällön puutteita. Pigeon tarkkailee etäisyys- ja sijaintiparametreja, kun taas Possum-algoritmin tarkoituksena on mahdollistaa paikallisten tulosten monipuolistaminen. (Patil et al., 2021)

Verkkohaun kieli voi vaikuttaa Googlen algoritmien tehokkuuteen. Alarifin ja muiden (2013) mukaan Googlen hakukoneen mustahattuhakukoneoptimoinnin vastaisien menetelmien vaikutukset ovat riittämättömiä arabinkielisissä sivustoissa. Menetelmien riittämätön testaaminen muissa kuin englanninkielessä aineistossa rohkaisee mustahattuhakukoneoptimoinnin harjoittajia kohdistamaan toimiansa muun ei-englanninkielisiin sivustoihin. (Alarifi et al., 2013)

9.7 Peliteorian soveltaminen negatiiviseen hakukoneoptimointiin

Verkkosivuston ylläpitäjät voivat hyödyntää *peliteoriaa* heikentääkseen negatiivisen hakukoneoptimoinnin vaikutuksia. Peliteorian avulla on mahdollista kuvailla tietoista ja tavoitteellista päätöksentekoprosessia, johon liittyy yksi tai useampi osapuoli, eli ”pelaaja.” Peliteorian avulla voidaan analysoida tilannetta, jossa kaksi tai useampi pelaaja tekevät itsenäisiä päätöksiä toisistaan riippumatta, mutta jossa lopputulos riippuu kaikkien pelaajien tekemistä päätöksistä. Jokainen pelaaja pyrkii tekemään päätöksiä, joista he saavat itselleen maksimaalisen hyödyn. *Nashin tasapaino* (Nash equilibrium) on tärkeä osa peliteoriaa. Se viittaa tilanteeseen, jossa yksikään pelaaja ei enää halua muuttaa päätöstään, ottaen huomioon toisen pelaajan tekemät päätökset. (Lynn et al., 2015)

Peliteoriaa sovellettaessa negatiiviseen hakukoneoptimointiin hyökkäävinä pelaajina ovat negatiivisen hakukoneoptimoinnin harjoittajat, ja puolustavina pelaajina yritykset, joiden verkkosivuille hyökkäykset kohdistuvat. Hyökkääjät ja puolustajat eivät tee yhteistyötä. Hyökkääjiä voi olla useita, ja he voivat tehdä yhteistyötä. Kaikki hyökkääjät katsotaan taidoiltaan tasavertaisiksi, eli jokaisesta hyökkääjän aikaansaamasta linkistä muodostuu uhka puolustajalle. Puolustajilla ei katsota olevan aikaisempaa historiaa huonolaatuisista linkeistä, eikä hakukoneyhtiöiden asettamista sanktioista. Jokainen manipulatiivinen linkki katsotaan toimivaksi, jos puolustaja ei ole hylännyt sitä. (Lynn et al., 2015)

Pelissä voi aiheutua puolustajalle kolme erilaista sanktiota hakukoneyhtiöiden toimesta: varoitus, hakutulossijoituksen alennus ja hakuindeksistä poistaminen. Pelissä jokainen siirto vaatii pelaajalta resursseja. Puolustajien tekemät ennaltaehkäisevät ja korjaavat toimenpiteet ovat lajiteltu kolmeen tasoon sen perusteella, kuinka paljon aikaa ja

resursseja ne vaativat ja kuinka suuri vaikutus sellaisella hyökkäyksellä on sijoittumiseen, johon korjaustoimi kohdistuu. Perustason toimenpiteisiin vaadittu aika ja taitotaso ovat alhaiset. Keskitason toimenpiteisiin kuluva aika ja taitotaso on korkeampi kuin perustasolla, ja toimenpiteet voivat vaatia parempaa teknistä taitoa tai jopa ammattitaitoa. Kehittyneen tason toimenpiteet vievät eniten aikaa ja todennäköisesti vaativat asiantuntijan apua. Verkkosivu saatetaan joutua vetämään pois väliaikaisesti. Vaatimustasojen mukaisia toimenpiteitä kuvaillaan taulukossa 5. (Lynn et al., 2015)

Puolustuksen taso	Kuvaus
Perustaso	Toimenpiteitä ovat: • Webmaster-työkalujen tarkkailu • Sivulle osoittavien linkkien tarkkailu • Poistamispyynnön antaminen linkin osoittavalle sivustolle ja epäilyttävien linkkien kieltäminen • Käyttäjien luomien huonolaatuisten linkkien sekä kommenttien tarkkailu ja poistaminen • Linkkien tai sivujen metatunnisteiden, tai eston sisältävän robots.txt muokkaaminen
Keskitaso	Sivun tarkkailu ja korjaustoimenpiteiden kohteena ovat: • Rakenteellinen merkintäkieli • CSS tyyli ja asettelu • Uudelleenohjaus • Lähdekoodi • Sisällönhallintajärjestelmä Toimenpiteisiin sisältyy myös uudelleenarviointipyynnön valmistelu ja lähettäminen.
Kehittynyt taso	Toimenpiteitä ovat : • Hakkeroidun sivun karanteeniin asettaminen • Aiheutuneiden vahinkojen arvioiminen • Haavoittuvuuksien tunnistaminen • Sivuston siistiminen ja ylläpitäminen • Sivuston siirtäminen uudelle verkkoisännöijälle • Uudelleenarviointipyynnön valmistelu ja lähettäminen.

Taulukko 5: Negatiiviselta hakukoneoptimoinnilta puolustettavalle sivustolle suoritettavat toimenpiteet vaatimustasoittain (Lynn et al., 2015)

Puolustajalle ja hyökkääjälle aiheutuvia kuluja kuvaillaan taulukossa 6.

Puolustajalle aiheutuvat kulut	Hyökkääjälle aiheutuvat kulut
• Prosessikulut, joihin lukeutuu perustason puolustukseen käytetty aika ja laskennalliset kulut • Materiaalikulut, joihin lukeutuu toimenpiteisiin käytetyt materiaalit ja raha • Vaihtoehtokustannukset, joihin lukeutuu alentuneesta hakutulossijoituksesta johtuva kävijäliikenteen lasku • Verkkosivun saavutettavuuden heikkenemisestä aiheutuneet kulut • Hakukoneyhtiön antaman luottamuksen ja auktoriteetin aleneminen	• Prosessikuluihin lukeutuu aika ja laskennalliset kulut, joita aiheuttavat negatiivisten profiilien, linkkien, verkkosivujen ja kommenttien luominen sekä kohdesivuston heikkouksien tunnistaminen ja hyväksikäyttö • Strategioiden toteuttamiseen kuluvat materiaaliset kulut ja raha

Taulukko 6: Negatiivisesta hakukoneoptimoinnista puolustajalle ja hyökkääjälle aiheutuvat kulut (Lynn et al., 2015)

Hyökkääjä hyötyy kohdesivun sijoituksen heikkenemisestä, mutta toimintaan liittyy riski paljastumisesta hakukoneille. Hyökkäysmenetelmiä kuvaillaan taulukossa 7. (Lynn et al., 2015)

Hyökkäysmenetelmä	Menetelmän kuvaus
Hakkeroitu verkkosivu	Hyökkääjän hakkeroima sivu, joka sisältää haittaohjelmia, epäluonnollisia linkkejä, epälaadukasta sisältöä tai muita mustahattumenetelmiä, kuten cloaking tai uudelleenohjaus. Sivun robots.txt voi olla muokattu torjumaan hakukoneita.
Sivulle osoitetut epäluonnolliset linkit	Sivulle osoittavat epäluonnolliset, keinotekoiset, petolliset tai manipuloivat linkit.
Mustahattuinen sisältö	Julkaisut tai profiilit, jotka ovat automaattisesti luotuja tai eivät ole autenttisen sivun käyttäjän luomia, tai sisältävät aiheeseen liittymättömiä sisältöä ja linkkejä.
Mustahattuinen verkkoisännöinti	Verkkoisännöintipalvelun sisältämistä sivustoista huomattava osa sisältää mustahattuisia sivuja.
Mustahattuisesta toiminnasta raportointi	Hakukoneyhtiölle ja muille tahoilla annettava raportti kohdesivuston harjoittamasta ohjesääntöjen vastaisesta toiminnasta.

Taulukko 7: Negatiivisessa hakukoneoptimoinnissa käytettyjä hyökkäysmenetelmiä (Lynn et al., 2015)

Pelaajien on otettava huomioon mahdolliset riskit sekä niiden toteutumisesta aiheutuvan vahingon voimakkuus. Hakukoneyhtiöt tunnistavat, että verkkosivuston ylläpitäjät eivät ole vastuussa kolmansien osapuolien niille osoittamista linkeistä, joten ne eivät välttämättä vaikuta negatiivisesti sivun sijoittumiseen. Linkit voivat mahdollisesti vaikuttaa sijoittumiseen myös positiivisesti. (Lynn et al., 2015)

Hyökkäys- ja puolustusmenetelmistä aiheutuvien kulujen, hyötyjen ja riskien ollessa tiedossa pelaajat voivat tutkia mahdollisten tilanteiden vaikutuksia ja pohtia strategiaa. Strategian laatimiseen voi hyödyntää taulukon 8 mukaista pohjaa eri tilanteiden aiheuttamien kulujen ja hyötyjen laskemiseksi. Vaikka kyseessä on hyvin yksinkertaistettu peliteoriaa soveltava malli, yritykset voivat hyödyntää peliteoriamallia riskien tunnistamiseen ja perustella sillä resurssien kohdistamista riskien vähentämiseen. (Lynn et al., 2015)

Hyökkäykset voidaan luokitella vahvoihin ja keskivahvoihin hyökkäyksiin sekä hyökkäämättömyyteen. Vahvojen ja keskivahvojen hyökkäyksien ero on niissä käytettyjen hyökkäysmenetelmien määrä. Mitä vahvempi hyökkäys, sitä enemmän hyökkääjältä vaaditaan resursseja. Puolustuksen tulee olla riittävä kohtuuttomien haittavaikutuksien välttämiseksi. Mitä kehittyneempi puolustus, sitä enemmän resursseja se vaatii puolustajalta. (Lynn et al., 2015)

	Perustason puolustus	Keskitason puolustus	Kehittyneen tason puolustus
Vahva hyökkäys			
Keskivahva hyökkäys			
Ei hyökkäystä			

Taulukko 8: Payoff matrix, jota voi hyödyntää strategian valintaan (Lynn et al., 2015)

Nashin tasapaino voi toteutua tilanteessa, jossa hyökkäystä ei toteuteta ja puolustajalla on perustason puolustus. Tilanne voi toteutua, jos hyökkääjän mahdollisesti saavuttamat hyödyt ovat hyvin rajoitteiset. Tilanteessa, jossa hyökkääjä käyttää vahvaa hyökkäystä ja puolustaja kehittynyttä puolustusta voi toteutua Nashin tasapaino, jos hakukoneyhtiön asettamat mahdolliset sanktiot ovat puolustajalle suuret ja hyökkääjälle pienet ja epätodennäköiset. Nashin tasapaino ei kuitenkaan toteudu, jos sanktiot ovat päinvastaisessa suhteessa. Näin ollen hakukoneyhtiöiden asettamat sanktiot voivat vaikuttaa pelaajien päätöksentekoon. Suuret sanktiot voivat kasvattaa hyökkääjien aktiivisuutta ja toisaalta puolustajien puolustusta. Hakukoneyhtiöt voivat vaikuttaa pelaajien toimintaan muuttamalla erityyppisten sanktioiden (varoituksen antaminen, hakutulossijoituksen alentaminen, hakuindeksistä poistaminen) suhteellista voimakkuutta tai muuttamalla

sanktioiden absoluuttista voimakkuutta. Jos hakukoneyhtiö valitsee rajoittavansa hyökkääjän toimintaa, se voi muuttaa sanktiokäytäntöjänsä esimerkiksi vähentämällä puolustajalle aiheutuvaa sanktiota, jotta hyökkäämättömyys olisi hyökkääjälle houkutteleva päätös. (Lynn et al., 2015)

10 Tulokset

Tässä luvussa esitetään tutkielman tulokset. Tutkimuskysymyksenä on mitä mustahattuhakukoneoptimoinnin menetelmiä ja vastatoimia tieteellisessä kirjallisuudessa on tutkittu. Tulokset on jaettu viiteen taulukkoon. Taulukossa 9 on sivun sisäiset mustahattumenetelmät, taulukoissa 10 ja 11 linkkiperustaiset mustahattumenetelmät, taulukossa 12 muut mustahattumenetelmät ja taulukossa 13 mustahattumenetelmien havainnointimenetelmät ja vastatoimet.

Sivun sisäiseille mustahattumenetelmille on ominaista huonolaatuinen, asiasisällöstä poikkeava sisältö, joka vähentää sivun käyttäjäkokemusta. Menetelmillä pyritään manipuloimaan sisältöperustaisesti hakutuloksia järjesteleviä hakukoneiden algoritmeja. Sisäiset mustahattumenetelmät kohdistuvat sivun sisältötekstiin ja HTML-sivun rakenteellisiin osiin. Sisäisten mustahattumenetelmien havainnointi perustuu niiden aiheuttamiin poikkeuksiin sivun sisällössä. Esimerkiksi sivun suuri sanamäärä ja sanojen toistuvuus voi viitata liialliseen avainsanojen käyttöön ja automaattisesti luotua sisältö voidaan havainnoida tutkimalla sisältötekstin ominaisuuksia, kuten luettavuutta ja sanaston rikkautta.

Taulukossa 9 kuvaillaan kirjallisuuskatsauksessa esiintyneitä sisäisiä mustahattumenetelmiä.

Sivun sisäiset mustahattumenetelmät	
Menetelmä	Kuvaus
Liiallinen avainsanojen käyttö	• Ylimääräisten, usein sivun aiheeseen liittymättömien avainsanojen tai sanayhdistelmien lisääminen sivulle. • Mahdollisena seurauksena sisällön luettavuuden heikkeneminen.
Sisällön piilottaminen	Sivun sisältö, joka tarkoitettu vain hakukoneiden ryömijöille. Toteutuskeinoina: • Huomaamaton fontti • Tekstin sisällyttäminen HTML-elementtiin, jonka sisältö ei näy sivulla • CSS hidden div
Automaattinen sattumanvaraisen sisällön luonti	• Automaattinen sattumanvarainen sisällön luonti sopivaa ohjelmistoa hyödyntäen.
Scraping	• Sisällön kopiointi toiselta verkkosivulta. • Eri lähteistä kopioitua tekstiä voidaan yhdistää tekstin muodostamiseksi (phrase stitching).
Text Spinning	• Tekstin kopioiminen ja muokkaaminen, millä pyritään estämään hakukoneita havaitsemasta sisällön plagiointia. • Voidaan toteuttaa manuaalisesti tai automaattisesti tehtävään soveltuvaa ohjelmistoa hyödyntäen. • Esimerkiksi The Best Spinner ja Spinbot mahdollistavat virkkeiden uudelleenjärjestämisen ja sanojen korvaamisen synonyymein, useiden versioiden luonnin yhdestä lähdetekstistä sekä luettavuuden ja kielioppitarkistuksen.

Taulukko 9: Sivun sisäiset mustahattumenetelmät

Linkkiperustaiset mustahattumenetelmät perustuvat linkkiperustaisesti hakutuloksia järjestelevien hakukoneiden algoritmien manipulointiin. Järjestelyalgoritmien lajitteluperusteita ovat kävijäliikenne, sivun osoittamien linkkien ja sivulle osoitettujen linkkien määrä ja laatu. Linkkiperustaisten mustahattumenetelmien havainnointi perustuu linkkien muodostamien verkostojen rakenteiden tutkimiseen sekä linkkien osoittamien sivustojen ominaisuuksien tutkimiseen. Taulukoissa 10 ja 11 kuvaillaan kirjallisuuskatsauksessa esiintyneitä linkkiperustaisia mustahattumenetelmiä.

Linkkiperustaiset mustahattumenetelmät 1/2	
Menetelmä	Kuvaus
Linkkien osoittaminen kohdesivulta	• Linkkien lisääminen sivulle manuaalisesti, tai esimerkiksi kopioimalla verkkosivujen osoitteita osoitekirjastosta.
Linkkifarmit	• Useiden sivustojen joukko, jotka osoittavat linkkejä usein toisillensa.
Spämmikommentit	• Linkkien lähettäminen blogien, foorumien tai sosiaalisen median palvelujen kommenttikenttiin. • Kommentin viesti ja kohdesivun aihe ovat usein epärelevantteja.
Spam blog	• Kohdesivulle ohjaavien blogisivujen ylläpitäminen. • Blogien päivitys saa aikaan hakukoneiden ryömijöiden ryömimään päivitettyjä sivuja (Blog-ping).
Blog-spam	• Ylläpidettävälle sivulle kohdistuvan linkin sisältävän spämmikommentin muokkaaminen lisäämällä kommenttiin omalle kohdesivulle osoittava linkki.
Verkkotunnuksien ostaminen linkkien osoittamiseksi	• Vanhentuneiden verkkotunnuksien hyödyntäminen linkkien osoittamiseen. • Spider-pool: Monimutkainen sivustorakenne, joka lisää sivustolla ryömintää.
Kohdesivuston kävijäliikennettä kasvattavat injektiot	• Haavoittuvan sivun hyödyntäminen kohdesivuston kävijäliikenteen nostattamiseen injektoimalla esimerkiksi uudelleenohjausominaisuus. • Injektio kohdistuu esimerkiksi HTML-rakenteeseen tai skriptiin.

Taulukko 10: Linkkiperustaiset mustahattumenetelmät 1/2

Linkkiperustaiset mustahattumenetelmät 2/2	
Menetelmä	Kuvaus
Bottiverkot ja manuaalinen click fraud	• Pay-per-click-mallin mukaisesti suoritettu automaattinen tai manuaalinen kävijäliikenteen kasvattaminen.
Kävijäliikennettä tarjoavat palvelut	• Palvelut tarjoavat vierailuja valitulle kohdesivustolle muilla sivustoilla vierailuja vastaan. • Osa palveluista on maksullisia, osa maksuttomia. • Palveluja on eri tasoisia, joista osa mahdollistaa automaattisen sivuilla vierailun.
Reputation manipulation services	• Maksulliset palvelut, jotka tarjoavat asiakkailleen hakutulossijoitusten kasvua ja näkyvyyden kasvua sosiaalisen median palveluissa. • Palveluita ovat esimerkiksi click fraud ja kohdesivustolle osoitetut linkit. • Palveluiden avaintekijät vastuussa suuresta osasta palveluiden tuottoa.
Private blog networks	• Tyypillisimmin blogisivustoista koostuva verkosto, joiden sivut osoittavat linkkejä asiakkaiden kohdesivustoille. • Usein asiakkaalle luodaan sivu, joka sisältää linkin kohdesivustolle sekä avainsanoja ja muuta sisältöä. • PBN-sivut usein luotu hyödyntäen sisällönhallintajärjestelmää, kuten WordPress.

Taulukko 11: Linkkiperustaiset mustahattumenetelmät 2/2

Taulukossa 12 kuvaillaan cloaking- ja uudelleenohjausmenetelmiä sekä niiden havainnointia.

Muut mustahattumenetelmät		
Menetelmä	Kuvaus	Havainnointi
Cloaking	Yhdestä verkkosivusta voidaan esittää eri sisältö selaimelle ja hakukoneiden ryömijöille. Menetelmiä vierailijan identiteetin tunnistamiseen ovat: • IP-cloaking • User-agent-cloaking • Referrer cloaking • Repeat-cloaking • JavaScript redirection cloaking	Sisällön ja linkkien tutkiminen vertailemalla selaimelle ja ryömijälle esitettäviä versioita.
Uudelleenohjaus	• Selaimen käyttäjä ohjataan automaattisesti toiselle sivulle. • Uudelleenohjaavat sivut voivat muodostaa vertaisverkon, jota pitkin vierailija ohjautuu useiden sivujen kautta kohdesivulle. Vertaisverkko heikentää uudelleenohjaavien sivujen havainnointia. • Uudelleenohjaava ominaisuus voidaan sisällyttää sivun skriptiin tai HTML-tunnisteeseen.	Uudelleenohjaavien ominaisuuksien havainnointi erityisesti sivun skriptistä ja skriptien vertailu alkuperäisiin js-lib-tiedostoihin.

Taulukko 12: Muut mustahattumenetelmät

Taulukossa 13 kuvaillaan kirjallisuuskatsauksessa esiintyneitä mustahattumenetelmien havainnointi- ja vastatoimia.

Mustahattumenetelmien havainnointi ja vastatoimet	
Menetelmä	Kuvaus
Hakukoneyhtiön antamat sanktiot	• Varoitus • Hakutulossijoituksen heikentäminen • Hakuindeksistä poistaminen • Mahdollisesti haitallista sisältävän sivuston merkitseminen hakutulossivulla
Google Zoo	Algoritmit, jotka • Tutkivat verkkosivujen ominaisuuksia ja pyrkivät pitämään hakutulokset laadukkaina ja relevantteina. • Pyrkivät heikentämään huonolaatuisten sivujen sijoittumista.
BadRank	• BadRank-arvo heikentää sivun arvoa hakutuloksissa. • Badrank-arvo kasvaa sivun osoittaessa linkkejä sivuille, joilla korkea BadRank -arvo.
TrustRank	• TrustRank-arvo kasvattaa sivun arvoa hakutuloksissa. • TrustRank perustuu olettamukseen, että hyvälaatuiset sivut harvoin osoittavat linkkejä huonolaatuisille sivustoille.
Itseoppivat algoritmit	• Luokittelevat sivuja mustahattuisiksi tai hyvälaatuisiksi sivujen sisältö- ja linkkiperustaisiin ominaisuuksiin perustuen. • Luovat säännönmukaisuuksia ja parantavat itse luokittelukykyään.
Verkkosivun alasajo oikeusteitse	• Verkkosivun toiminta voidaan pyrkiä estämään oikeusteitse.
Peliteoria	• Tukee tietoista ja tavoitteellista päätöksentekoprosessia. • Peliteoriaa voidaan soveltaa negatiivisen hakukoneoptimoinnin vaikutuksien vähentämiseksi.

Taulukko 13: Mustahattumenetelmien havainnointi- ja vastatoimet

11 Keskustelua

Osa kirjallisuuskatsauksessa tutkitusta aineistosta on julkaistu useita vuosia ennen tutkielman tekoa, minkä vuoksi tutkielmassa kuvaillaan jo pitkään olemassa olleita mustahattumenetelmiä. Jatkotutkimuksen kannalta on mielenkiintoista keskittyä mahdollisiin tuoreempiin menetelmiin. Mustahattuhakukoneoptimoinnin vastatoimiin hyödynnettäviä algoritmia on tutkittu useissa tutkimuksissa. Mustahattumenetelmistä liiallinen avainsanojen käyttö ja cloaking-menetelmä toistuvat usein kirjallisuudessa. Tutkielman kirjallisuuskatsauksen otanta on alle viisikymmentä lähdettä, joten kirjallisuuskatsauksen ulkopuolelle on voinut jäädä useita tutkimattomia menetelmiä. Otannan laajentamisella ja muiden tutkimusmenetelmien hyödyntämisellä jatkotutkimuksessa voi saada mustahattuhakukoneoptimoinnin menetelmistä ja vastatoimista laajemmin informaatiota, informaation ollessa myös päivitetympää. Jatkotutkimuksessa voi syventyä tarkemmin yksittäiseen mustahattumentelmään, kuten cloaking-menetelmään. Lisäksi mustahattuhakukoneoptimoinnin vaikutuksista voi saada konkreettista informaatiota julkaisemalla eri mustahattumenetelmiä hyödyntäviä verkkosivuja verkossa ja seuraamalla, miten hakukoneet käsittelevät verkkosivua.

Osa tässä tutkielmassa tutkituista aineistoista on julkaistu yli kymmenen vuotta sitten. On mahdollista, että spämmiin liittyvä tutkimus keskittyy nykyisin enemmän muihin yhteyksiin, kuin hakukoneoptimointiin. Jatkotutkimuksessa spämmiä voi tutkia esimerkiksi sosiaalisen median yhteydessä. ChatGPT:n ollessa yleisen keskustelun aiheena yksi mielenkiintoinen tutkimusaihe on tekoälyn vaikutus mustahattuhakukoneoptimointiin ja muuhun spämmiin. Tekoäly mahdollistaa automaattisen sisällön luomisen, joka ei vaikuta lukijalle välttämättä keinotekoiselta. Hakukoneoptimoinnin yhteydessä se voi tarkoittaa esimerkiksi verkkosivun sisällön toteuttamista valittujen avainsanojen ympärille ilman kopiointia. Tässä tutkielmassa kuvailtiin, miten spinning-ohjelmiston tuottamaa tekstiä voidaan havainnoida perustuen tekstin tiettyihin ominaisuuksiin. Tekoälyn kehittyessä onkin mielenkiintoista, miten sen luomaa tekstiä voidaan tunnistaa. Mustahattumenetelmien kehittyessä vastatoimienkin pitää kehittyä.

12 Yhteenveto

Tämän tutkielman tavoitteena oli selvittää kirjallisuuskatsauksen pohjalta, mitä mustahattuhakukoneoptimoinnin menetelmiä ja vastatoimia tieteellisessä kirjallisuudessa on tutkittu. Hakukoneoptimoinnilla pyritään kasvattamaan verkkosivun sijoitusta verkkohakukoneiden tulossivuilla. Hakukoneyhtiöiden ohjesääntöjen mukaisia hakukoneoptimointimenetelmiä kutsutaan valkohattuiseksi hakukoneoptimoinniksi ja ohjesääntöjen vastaisia menetelmiä kutsutaan mustahattuiseksi hakukoneoptimoinniksi. Harmaahattuiseksi hakukoneoptimoinniksi kutsutaan sellaista hakukoneoptimointia, joka ei ole selvästi mustahattuista tai valkohattuista. Mustahattumenetelmin pyritään manipuloimaan

hakutuloksia vaikuttamalla hakukoneiden järjestelyalgoritmeihin, eikä verkkosivulle pyritä luomaan laadukasta ja relevanttia sisältöä. Mustahattuhakukoneoptimointi aiheuttaa epälaadukkaan sisällön korostumisen hakukoneiden tulossivuilla ja kuluttaa hakukoneiden resursseja. Lisäksi mustahattumenetelmiä hyödyntävät verkkosivut voivat olla yhteydessä haitallisen verkkosisällön levittämiseen ja aiheuttaa esimerkiksi haittaohjelmien leviämisen käyttäjien laitteille.

Mustahattumenetelmät voidaan jakaa sivun sisäisiin ja linkkiperustaisiin menetelmiin sekä tehostaviin menetelmiin ja piilotusmenetelmiin. Tehostavat menetelmät pyrkivät nostamaan verkkosivun hakutulossijoitusta ja piilotusmenetelmät ehkäisemään tehostusmenetelmien käyttöä tulemasta havainnoiduksi. Sivun sisäiset mustahattumenetelmät kohdistuvat verkkosivun sisältöön ja rakenteeseen. Linkkiperustaiset mustahattumenetelmät kohdistuvat sivun osoittamiin ja sivulle osoitettuihin linkkeihin. Monimutkaiset linkkirakenteet heikentävät mustahattumenetelmien havainnointia. Muita mustahattumenetelmiä ovat cloaking ja uudelleenohjaus. Mustahattumenetelmiä voidaan hyödyntää myös negatiiviseen hakukoneoptimointiin.

Mustahattumenetelmiä hyödyntäviä sivuja havainnoidaan niiden ominaisuuksiin perustuen. Automaattiset havainnointimenetelmät tehostavat havainnointia ja tekevät havainnoinnista tarkempaa. Hakukoneyhtiöt hyödyntävät algoritmeja havainnoidakseen verkkosivujen sisäisiä ja linkkiperustaisia ominaisuuksia mustahattumenetelmien havainnoimiseksi. Havaitessaan mustahattumenetelmien käyttöä hakukoneyhtiöt voivat antaa verkkosivun ylläpidolle varoituksen, heikentää verkkosivun sijoitusta hakutuloksissa, tai poistaa verkkosivun hakuindeksistä. Mustahattumenetelmiä hyödyntävien verkkosivujen alasajo oikeusteitse on tehotonta, koska mustahattumenetelmien harjoittaminen voidaan siirtää toiseen sijaintiin. Peliteoriaa voidaan soveltaa negatiivista hakukoneoptimointia ennaltaehkäisevien ja korjaavien toimenpiteiden suunnitteluun.

Jatkotutkimuksessa voi laajentaa tutkittavien menetelmien kirjoa ja kiinnittää huomion tuoreempiin menetelmiin. Toisaalta, tutkittavien menetelmien määrää voi vähentää ja syventyä yhteen menetelmään. Spämmiä voi tutkia muissa yhteyksissä, kuten sosiaalisessa mediassa. Tekoälyn vaikutus mustahattumenetelmiin on mielenkiintoinen tutkimusaihe.

13 Viiteluettelo

Alarifi, A., Alsaleh, M., Al-Salman, A., Alswayed, A., & Alkhaledi, A. (2013). Google Penguin: Evasion in Non-English Languages and a New Classifier. 2013 12th International Conference on Machine Learning and Applications, 2, 274–280. https://doi.org/10.1109/ICMLA.2013.135

Croft, W. B., Metzler, D., & Strohman, T. (2009). Search engines: information retrieval in practice. Pearson Addison-Wesley.

Deng, J., Chen, H., & Sun, J. (2013). Uncovering Cloaking Web Pages with Hybrid Detection Approaches. 2013 International Symposium on Computational and Business Intelligence, 291–296. https://doi.org/10.1109/ISCBI.2013.65

Du, K., Yang, H., Li, Z., Duan, H., & Zhang, K. (2016). The ever-changing labyrinth: A large-scale analysis of wildcard DNS powered blackhat SEO. Proceedings of the 25th USENIX Security Symposium, 245–262.

Duk, S., Bjelobrk, D., & Carapina, M. (2013). SEO in e-commerce: Balancing between white and black hat methods. 2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013 - Proceedings, 390–395.

Farooqi, S., Jourjon, G., Ikram, M., Kaafar, M. A., De Cristofaro, E., Shafiq, Z., Friedman, A., & Zaffar, F. (2017). Characterizing key stakeholders in an online black-hat marketplace. eCrime Researchers Summit, eCrime, 17–27. https://doi.org/10.1109/ECRIME.2017.7945050

Fox, V. (2008). Black hat or white hat? Ethical SEO tactics are more straightforward than you might think. But a solid understanding of guidelines will help you stay on the straight-and-narrow. Audience Development, 23(10), 36–.

Gandour, A., & Regolini, A. (2011). Web site search engine optimization: a case study of Fragfornet. Library Hi Tech News, 28(6), 6–13. https://doi.org/10.1108/07419051111173874

Ghiam, S. (2012). A Survey on Web Spam Detection Methods: Taxonomy. International Journal of Network Security & Its Applications, 4(5), 119–134. https://doi.org/10.5121/ijnsa.2012.4510

Ghosh, S., & Desarkar, M. S. (2018). Class Specific TF-IDF Boosting for Short-text Classification: Application to Short-texts Generated during Disasters. The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018, 1629–1637. https://doi.org/10.1145/3184558.3191621

Gudivada, V. N., Rao, D., & Paris, J. (2015). Understanding Search-Engine Optimization. Computer (Long Beach, Calif.), 48(10), 43–52. https://doi.org/10.1109/MC.2015.297

Gyongyi, Z., & Garcia-Molina, H. (2005). Web spam taxonomy. Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, AIRWeb 2005 - Held in Conjunction with the 14th International World Wide Web Conference, 39–47.

Javed, M., Herley, C., Peinado, M., & Paxson, V. (2015). Measurement and analysis of traffic exchange services. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, 2015-, 1–12. https://doi.org/10.1145/2815675.2815708

Jha, T., & Saraswat, S. (2018). Selecting the Best Approach for Website Optimization. 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), 555–559. https://doi.org/10.1109/ICGCIoT.2018.8753107

Killoran, J. B. (2013). How to Use Search Engine Optimization Techniques to Increase Website Visibility. IEEE Transactions on Professional Communication, 56(1), 50–66. https://doi.org/10.1109/TPC.2012.2237255

Kumar, S., Xiaoying Gao, Welch, I., & Mansoori, M. (2016). A Machine Learning Based Web Spam Filtering Approach. 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), 2016-, 973–980. https://doi.org/10.1109/AINA.2016.177

Levene, M. (2010). An introduction to search engines and web navigation (2nd ed.). Wiley.

Li, H. (2014). Web spam detection based on improved tri-training. PIC 2014 - Proceedings of 2014 IEEE International Conference on Progress in Informatics and Computing, 61–65. https://doi.org/10.1109/PIC.2014.6972296

Li, Z. Alrwais, S., Wang, X., & Alowaisheq, E. (2014). Hunting the Red Fox Online: Understanding and Detection of Mass Redirect-Script Injections. 2014 IEEE Symposium on Security and Privacy, 3–18. https://doi.org/10.1109/SP.2014.8

Liu, J., Su, Y., Lv, S., & Huang, C. (2020). Detecting Web Spam Based on Novel Features from Web Page Source Code. Security and Communication Networks, 2020, 1–14. https://doi.org/10.1155/2020/6662166

Lynn, T., Brady, M., & Masevic, I. (2015). A risk assessment method for negative SEO attacks using a game theoretic approach. 2015 IEEE International Professional Communication Conference (IPCC), 2015-, 1–12. https://doi.org/10.1109/IPCC.2015.7235795

Ma, X. (2018). Research on Black Hat SEO Behaviour Measurement. 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 1041–1045. https://doi.org/10.1109/IAEAC.2018.8577831

Malaga, R. (2008). Worst practices in search engine optimization. Communications of the ACM, 51(12), 147–150. https://doi.org/10.1145/1409360.1409388

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). Detecting spam web pages through content analysis. Proceedings of the 15th International Conference on World Wide Web, 83–92. https://doi.org/10.1145/1135777.1135794

Patil, A., Pamnani, J., & Pawade, D. (2021). Comparative Study Of Google Search Engine Optimization Algorithms: Panda, Penguin and Hummingbird. 2021 6th International Conference for Convergence in Technology (I2CT), 1–5. https://doi.org/10.1109/I2CT51068.2021.9418074

Pevtsov, S., & Volkov, S. (2013). Russian web spam evolution: yandex experience. Proceedings of the 22nd International Conference on World Wide Web, 1137–1140. https://doi.org/10.1145/2487788.2488135

Poulimenou, S., Papavlasopoulos, S., Kapidakis, S., & Poulos, M. (2016). Towards to Vector Plain Model. ACM International Conference Proceeding Series, 29-, 1–2. https://doi.org/10.1145/2910674.2910693

Raiber, F., Collins-Thompson, K., & Kurland, O. (2013). Shame to be sham: addressing content-based grey hat search engine optimization. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1013–1016. https://doi.org/10.1145/2484028.2484135

Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press.

Roslina, A., & Shahirah, M. (2019). Implementing white hat search engine technique in e-business website. ACM International Conference Proceeding Series, 311–314. https://doi.org/10.1145/3306500.3306533

Salton, G., & McGill, M. J. (1987). Introduction to modern information retrieval (3rd Printing). McGraw-Hill.

Shahid, U., Farooqi, S., Ahmad, R., Shafiq, Z., Srinivasan, P., & Zaffar, F. (2017). Accurate detection of automatically spun content via stylometric analysis. Proceedings - IEEE International Conference on Data Mining, ICDM, 2017-, 425–434. https://doi.org/10.1109/ICDM.2017.52

Shahzad, A., Nawi, N. M., Rehman, M. Z., & Khan, A. (2021). An Improved Framework for Content- and Link-Based Web-Spam Detection: A Combined Approach. Complexity (New York, N.Y.), 2021, 1–18. https://doi.org/10.1155/2021/6625739

Somani, A., & Suman, U. (2011). Counter measures against evolving search engine spamming techniques. 2011 3rd International Conference on Electronics Computer Technology, 6, 214–217. https://doi.org/10.1109/ICECTECH.2011.5942084

Svore, K., Wu, Q., Burges, C., & Raman, A. (2007). Improving web spam classification using rank-time features. ACM International Conference Proceeding Series, 215, 9–16. https://doi.org/10.1145/1244408.1244411

Taweesiriwate, A., Manaskasemsak, B., & Rungsawang, A. (2012). Web Spam Detection Using Link-Based Ant Colony Optimization. 2012 IEEE 26th International Conference on Advanced Information Networking and Applications, 868–873. https://doi.org/10.1109/AINA.2012.118

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. The Journal of Artificial Intelligence Research, 37, 141–188. https://doi.org/10.1613/jair.2934

Van Goethem, T., Miramirkhani, N., Joosen, W., & Nikiforakis, N. (2019). Purchased Fame: Exploring the Ecosystem of Private Blog Networks. AsiaCCS 2019 - Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, 366–378. https://doi.org/10.1145/3321705.3329830

Varsha, Grover, P. S., & Ahuja, L. (2021). An Overview of Search Engine Optimization. 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1–6. https://doi.org/10.1109/ICRITO51393.2021.9596287

Visser, E. B., & Weideman, M. (2011). An empirical study on website usability elements and how they affect search engine optimisation. South African Journal of Information Management, 13(1), C1–e9. https://doi.org/10.4102/sajim.v13i1.428

Wang, D., Savage, S., & Voelker, G. (2011). Cloak and dagger: dynamics of web search cloaking. Proceedings of the 18th ACM Conference on Computer and Communications Security, 477–490. https://doi.org/10.1145/2046707.2046763

Wang, D., Der, M., Karami, M., Saul, L., McCoy, D., Savage, S., & Voelker, G. (2014). Search + Seizure: The Effectiveness of Interventions on SEO Campaigns. Proceedings of the 2014 Conference on Internet Measurement Conference, 359–372. https://doi.org/10.1145/2663716.2663738

Weideman, M. (2009). Website Visibility: The Theory and Practice of Improving Rankings. https://doi.org/10.1533/9781780631790

Wu, B., & Davison, B. (2005). Identifying link farm spam pages. 14th International World Wide Web Conference, WWW2005, 820–829. https://doi.org/10.1145/1062745.1062762

Wynne, P. (2012). Pimp my site the DIY guide to SEO, search marketing, social media and online PR (1st edition). Capstone Publishing.

Yang, R., Liu, J., Gu, L., & Chen, Y. (2020). Search & Catch: Detecting Promotion Infection in the Underground through Search Engines. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 1566–1571. https://doi.org/10.1109/TrustCom50675.2020.00216

Zhang, J., & Dimitroff, A. (2005). The impact of webpage content characteristics on webpage visibility in search engine results (Part I). Information Processing & Management, 41(3), 665–690. https://doi.org/10.1016/j.ipm.2003.12.001

Zuze, H., & Weideman, M. (2013). Keyword stuffing and the big three search engines. Online Information Review, 37(2), 268–286. https://doi.org/10.1108/OIR-11-2011-0193