Pentti Koskela

DATAN ANALYSOINTI VERKKOKAUPPOJEN SUOSITTELUJÄRJESTELMISSÄ BIG DATAN KONTEKSTISSA

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2015

TIIVISTELMÄ

Koskela, Pentti
Datan analysointi verkkokauppojen suosittelujärjestelmissä big datan kontekstissa
Jyväskylä: Jyväskylän yliopisto, 2015, 39 s.
Tietojärjestelmätiede, kandidaatin tutkielma
Ohjaaja: Mazhelis, Oleksiy

Tämä tutkielma keskittyy suosittelujärjestelmien yleiseen toimintaperiaatteeseen ja niiden tarjoamiin hyötyihin. Suosittelujärjestelmien toiminta täyttää Big datalle ominaiset piirteet, mistä syystä asiaa lähestytään Big datan analysointina. Tarkoituksena on antaa lukijalle yleiskuva suosittelujärjestelmistä, niiden toimintaperiaatteesta ja käyttötarkoituksesta. Tutkielmassa käsitellään suosittelujärjestelmistä yhteisöllistä, yhteistoimintapohjaista, demografista ja tietämyspohjaista suodatusta sekä hybridejä variaatioita. Lisäksi kerrotaan suosittelujärjestelmien tarjoamista hyödyistä nimenomaan palveluntarjoajan näkökulmasta.

Tutkielma on toteutettu kirjallisuuskatsauksena. Lähdeaineisto koostuu pääosin tieteellisistä artikkeleista ja muutamasta aiheeseen liittyvästä kirjasta, jotka on julkaistu pääosin 2000-luvulla. Yleisesti tiedossa olevissa asioissa on viitattu kaupallisiin lähteisiin ja blogikirjoituksiin. Lähdeaineistoa valitessa yhtenä prioriteettina oli julkaisuajankohta, sillä suosittelujärjestelmät ovat melko uusi ja nopeasti kehittyvä ala.

Tämän tutkielman luettuaan lukijalla on yleiskäsitys suosituimpien suodatustekniikoiden toimintaperiaatteesta. Suosittelujärjestelmien tarjoamat hyödyt ja huomionarvoiset asiat suosittelujärjestelmien käyttöönotossa tulevat myös esille. Verkkokaupalle ominainen pitkä häntä-ilmiö käsitellään omana lukunaan.

Asiasanat: suosittelujärjestelmät, big data, verkkoliiketoiminta, yhteisöllinen suodatus, sisältöpohjainen suodatus, pitkä häntä

ABSTRACT

Koskela, Pentti
Data analysis in e-commerce recommendation systems in big data context
Jyväskylä: University of Jyväskylä, 2015, 39 p.
Information Systems, Bachelor's Thesis
Supervisor: Mazhelis, Oleksiy

This research paper's main focus is on general function of recommender systems and what benefits they will give. Recommender systems function fulfills the principles of big data and therefore this area is approached as a big data context. The idea is to provide general view about recommender systems, how they function and for what purpose they are designed. This paper focuses on collaborative filtering, content-based filtering, demographic filtering, knowledge-based filtering and hybrid variations. In addition to that, possible commercial benefits are discussed in a service provider point-of-view.

This research is made as a literature review. References consists of scientific articles and a few area related books, most of these published in 2000-era. Commercial sources and blogs have been used in a few relatively commonly known issues. One of the priorities in selecting references was publishing date because recommender systems is fast-developing area of interest.

After reading this research paper reader should have general understanding about most common recommendation filtering techniques. Commercial benefits and what to take in consideration when applying recommender system in use are also pointed out. The long tail, phenomenon occurring in e-commerce, is also addressed as its own chapter.

Keywords: recommender systems, big data, e-commerce, collaborative filtering, content-based filtering, the long tail

KUVIOT

TAULUKOT

SISÄLLYS

1 Johdanto

Internetissä toimivien verkkokauppojen kilpailuetuna ei ole enää sijainti, vaan suuremmassa merkityksessä on tarjolla olevien tuotteiden määrä ja hinta. Yhteistä molemmille kaupoille on erinomaisen asiakaspalvelun tarjoaminen. Verkkokaupoissa tuotteita voi olla satoja tuhansia, jopa miljoonia. Asioiminen näin suuren valikoiman kaupassa ja sen oikean tuotteen löytäminen voi olla erittäin hidasta ja turhauttavaa. Kilpaileva kauppa on vain muutaman hiiren klikkauksen päässä, joten asiakaskokemuksesta on saatava mahdollisimman miellyttävä mahdollisimman nopeasti.

Tähän ongelmaan on vastattu kehittelemällä erilaisia suosittelujärjestelmiä. Käytännössä jokaisella suuremmalla verkkokaupalla on jokin variaatio suosittelujärjestelmästä apunaan. Kaupallisessa mielessä suosittelujärjestelmät ovat pakollinen käyttöönotto verkkokaupalle, sillä muuten kilpailevat verkkokaupat suoriutuvat paremmin asiakaspalvelussa. Asiakkaan mielenkiinto verkkokauppaa kohtaan on saatava välittömästi, mikä asettaa haasteita suosittelujärjestelmien toteutuksessa.

Suosittelujärjestelmiä kohtaa nykypäivänä lähes kaikkialla internetissä, oli kyseessä sitten matkailusivusto, musiikki-/videopalvelu tai verkkokauppa. Tämä tutkielma perustuu suosittelujärjestelmien käyttöön nimenomaan verkkokauppojen kontekstissa. Näin ollen tässä tutkielmassa suosittelujärjestelmää käyttävästä henkilöstä puhutaan nimellä käyttäjä tai asiakas. Suositeltava kohde on nimeltään tuote. Eri suosittelujärjestelmätyypeistä käytetään lyhenteitä, jotka muodostuvat niiden englanninkielisistä nimistä. Lyhenteiden takia tekstissä on vähemmän toistoa ja sen lukeminen on sujuvampaa. Englanninkieliset lyhenteet ovat rinnan muiden samaa aihetta käsittelevien tutkielmien kanssa.

Suosittelujärjestelmät on uudehko tutkimusalue ja sitä on käsitelty tieteenalana vasta 1990-luvun puolivälistä lähtien (Anand & Mobasher, 2005) Viime vuosina mielenkiinto suosittelujärjestelmiä kohtaan on kasvanut huomattavasti. Yksi syy tähän on useiden erittäin suosittujen verkkopalveluiden toiminnan riippuvuus suosittelujärjestelmistä, kuten Amazon.com, Youtube, Netflix, ja IMDB. Suosion kasvusta kertoo myös se, että vuodesta 2007 lähtien on järjestetty vuosittainen ACM Recommender Systems – konferenssitapahtuma.

Suosittelujärjestelmien kehittäminen on haastavaa. Ricci, Rokach, Shapira ja Kantor (2011, 6) toteavat kirjassaan *Recommender Systems Handbook* seuraavasti:

> Suosittelujärjestelmien kehittäminen on monialainen prosessi, mihin osallistuu asiantuntijoita usealta eri osaamisalueelta, kuten tekoälystä, ihminen – kone vuorovaikutuksesta, informaatioteknologiasta, tiedon louhinnasta, tilastotieteestä, käyttöliittymistä, päätöksenteon tukijärjestelmistä, markkinoinnista ja kuluttajakäyttäytymisestä.

Alan nuori ikä asettaa haasteita lähdemateriaalin löytymisessä, mikä näkyy aiheesta riippuen vajavaisena lähteiden määränä. Erityisesti suosittelujärjestelmien kaupallisia etuja käsitellään vain yleisellä tasolla ilman tarkempia lukuja esimerkiksi myynnin kasvusta. Koska suuret verkkokaupat eivät näe edukseen julkaista yksityiskohtaisia etuja heidän suosittelujärjestelmistään, jäävät tämänkin tutkielman käsittelemät edut alkuperäistä yleisemmälle tasolle.

Tämä tutkielma pyrkii löytämään vastauksen seuraaviin tutkimuskysymyksiin:

- Mitä työkaluja käytetään analysoinnin tukena?
- Mitä hyötyjä verkkokauppojen suosittelujärjestelmät tarjoavat palvelun tarjoajalle?

Suurin osa aiemmasta tutkimusmateriaalista keskittyy suosittelujärjestelmien tekniseen toteutukseen. Odotetusti monet näistä tutkielmista ovat uutta teoriaa luovia esittäen uuden toimintatavan esimerkiksi algoritmin muodossa. Nämä tutkielmat ovat syvästi teknisiä näkemyksiä ja ne keskittyvät vain käsiteltävään suosittelujärjestelmämalliin. Pieni osa aiemmista tutkimuksista keskittyy kaupalliseen aspektiin suosittelujärjestelmissä, joissa on otettu tapaustutkimuksen omaisesti tarkempaan analyysiin yksi tai useampi suosittelujärjestelmiä käyttävä yritys.

Mikä on tehokkain tapa saada tietoa asiakkaan mieltymyksistä mielenkiintoisen tarjonnan aikaansaamiseksi? Millä tavalla tietoja käyttäjistä saadaan? Muun muassa näihin kysymyksiin löytyy vastaus tästä tutkielmasta. Tämän tutkimuksen tarkoitus on yhdistää sekä teknistä aspektia että suosittelujärjestelmien tarjoamia hyötyjä aiempien tutkimuksien pohjalta. Syvempi tekninen käsitteleminen ei ole mielekästä kanditaatintutkielman laajuisessa työssä. Aihealue muuttuisi myös tällöin enemmän tietotekniikka- tai matematiikkakeskeiseksi.

Etuja käsittelevässä luvussa huomioitavaa on pitkä häntä-ilmiön käsitteleminen omana alalukunaan. Tätä ei ole käsitelty aiemmassa kirjallisuudessa suosittelujärjestelmien yhteydessä kuin pintapuolisesti, jos ollenkaan. Toisena alalukuna ovat tavat näyttää suosittelujärjestelmiä verkkokaupan sivustolla asiakkaalle. Hieman yllättäen monet tavat ovat yhteydessä perinteisessä kaupankäynnissä oleviin menetelmiin. Kaikki tavat eivät sovellu kaikkiin käyttö-

tarkoituksiin, joten suosittelujärjestelmää harkitsevan on otettava tämä huomioon suunnittelussa ja käyttöönotossa.

Havaitsin aihetta käsitteleviä tutkimuksia lukiessani, että big dataa ei ole mainittu juuri koskaan, jos tutkielman aihe on ollut suosittelujärjestelmät. Mikäli tutkimus on käsitellyt big dataa, on siinä ollut viittauksia suosittelujärjestelmiin. Big datasta ja sen yhteydestä suosittelujärjestelmiin on kirjoitettu tässä tutkielmassa luvussa 2.

Luku 3 käsittää suosituimmat suosittelujärjestelmätyypit jakaen suodatustekniikat viiteen kategoriaan: yhteisölliset, sisältöpohjaiset, tietämyspohjaiset, demografiset ja hybridit. Hybridit suosittelujärjestelmät ovat nimensä mukaisesti yhdistelmiä kahdesta tai useammasta suodatustekniikasta. Neljä ensimmäistä suodatustekniikkaa ja niiden tärkeimmät komponentit näytetään luvun alkupuolella, josta saa hyvän yleiskuvan niiden yhtäläisyyksistä. Yhteisöllisestä ja sisältöpohjaisesta suosittelujärjestelmästä kerrotaan eniten, sillä ne edustavat sekä vanhinta että suosituinta suodatustekniikkaa. Niiden toimintaperiaatteen kuvaaminen vaatii myös enemmän sisältöä verrattuna demografiseen ja tietämyspohjaiseen suodatukseen. Hybridien suodatustekniikoiden jaossa on useita näkemyksiä. Päädyin Burken (2007) esittämään malliin, missä hybridit tekniikat on jaettu seitsemään osaan. Tämä näkyy kyseisessä luvussa lähteiden vähäisenä määränä, sillä muissa tutkimuksissa aihetta on lähestytty eri tavalla.

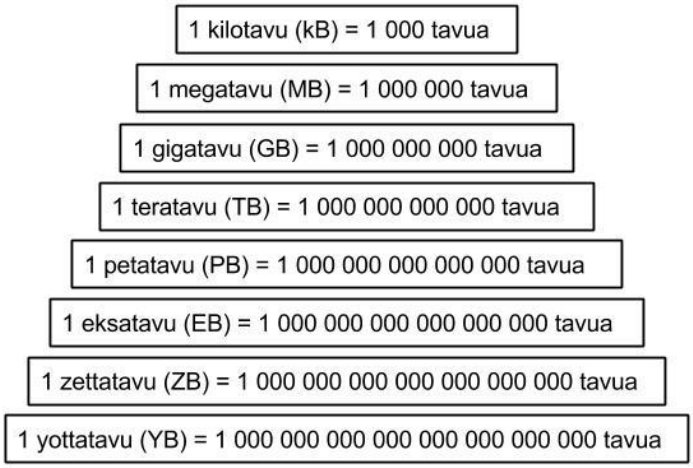
Luvussa 4 kerrotaan suosittelujärjestelmien tarjoamista hyödyistä palveluntarjoajan näkökulmasta. Sen lisäksi kerrotaan, mitä on alustavasti otettava huomioon suosittelujärjestelmän suunnittelussa ja käyttöönotossa. Myös erilaiset tavat näyttää suodatettu sisältö käyttäjälle käsitellään tässä luvussa. Luku 4.1. selittää pitkä häntä-ilmiön määritelmän ja taustaa. Alaluku 4.2. kuvaa, mitä asioita on otettava huomioon suosittelujärjestelmän käyttöönotossa.

Luku 5 on tutkielman päättävä luku, missä muodostetaan tutkielmassa saavutetut tulokset sekä käsitellään jatkotutkimusaiheita. Tutkielman lopussa on esimerkkejä eri suosittelujärjestelmistä liitetiedostoina. Esimerkit on poimittu Netflix-suoratoistopalvelusta. Netflixin liiketoimintamalli ei edusta verkkokauppaa, mutta se käy hyvästä esimerkistä, sillä sen suosittelujärjestelmäkomponentit sisältävät kaikki tässä tutkielmassa esitellyt suodatustavat.

2 Big Data ja suosittelujärjestelmät

Tiedon määrä on kasvanut räjähdysmäisellä vauhdilla viimeisen vuosikymmenen aikana. Sivilisaation alusta vuoteen 2003 dataa on kerätty noin viiden eksatavun verran. Googlen pääjohtajan Eric Schmidtin mukaan saman verran dataa syntyy nykyään joka toinen päivä (Techcrunch, 2010.). Vuonna 2012 dataa oli kerätty 2,72 zettatavua ja luku tulee tuplaantumaan noin kahden vuoden välein (Intel, 2013).

Termillä Big data viitataan tähän valtavaan, hajautuneeseen, nopeasti kasvavaan määrään dataa. Datan eksponentiaalisen kasvun vaikuttajina voidaan pitää muun muassa digitalisaatiota, teollisuuden internetiä (engl. Internet of Things, IoT), älypuhelimia ja erilaisia sensorilaitteita (Intel, 2013.). Dataa kerääntyy enemmän mitä sitä pystytään tehokkaasti käsittelemään. Big datan yhteydessä puhutaan erittäin suurista tietomääristä, joita mitataan teratavuina, petatavuina ja eksatavuina. Kuvio 1 näyttää nämä määreet perspektiivissä.



KUVIO 1 Tallennuskapasiteetin mittayksiköt (Intel, 2003)

Big datan määritteleminen on vaikeaa, eikä siihen ole löytynyt konsensusta. Big datalla on kuitenkin useita yhteneviä ja toistuvia piirteitä. Franks (2012) mainitsee viisi asiaa, mitkä yleensä ilmenevät Big datan yhteydessä:

- Big data on jonkun laitteen automaattisesti tuottamaa dataa.
- Tietoa kerätään uudenlaisista lähteistä.
- Monet tiedonlähteet eivät ole suunniteltu tiedon keräämistä varten.
- Suuri osa kerätystä tiedosta on hyödytöntä ja turhaa dataa.
- Monet tietolähteet ovat strukturoimattomia tai semi-strukturoituja.

Usein Big datan yhteydessä puhutaan kolmen V:n määritelmästä: volume (suom. volyymi, tilavuus), velocity (suom. nopeus), variety (suom. vaihtelu, varieteetti) Kolmen V:n määritelmä on laatinut alun perin Doug Laney blogissaan otsikolla 3D Data Management: Controlling Data Volume, Velocity, and Variety. Hän viittasi näillä nimenomaan datamäärän lisääntymiseen sähköisessä kaupankäynnissä (Laney, 2001.).

Volume
Volyymilla viitataan Big datalle ominaiseen suuren datan keräämiseen ja sen käsittelemiseen. Suuresta datamäärästä on etua analyysejä ja ennustuksia tehdessä. Haittapuolena on datan säilöminen ja tehokas prosessointi sekä suuri ”turhan” datan määrä.

Velocity
Big datassa tietoa kerätään usein reaaliajassa, joka vaatii tehokasta ja nopeaa tiedonkäsittelyä. Asiakkaan verkkokaupan sivustolla tekemät valinnat ja toiminnot siirtyvät reaaliajassa analysoitavaksi, esimerkiksi suosittelujärjestelmiä varten. Tieto voi myös olla ns. liikkuvaa (engl. streaming data), esimerkiksi tiedon analysointi median suoratoistopalveluissa.

Variety
Dataa kerätään lukuisista eri lähteistä ja useasta eri formaatista, mikä aiheuttaa suurta vaihtelua, varieteettia. Data on myös hajautunutta ja usein sellaisessa muodossa, mitä perinteiset data-analyysityökalut eivät osaa käsitellä (esim. kuvan ja videon metatiedot, erilainen sensorien tuottama data). Monet datalähteet ovat sen tyyppisiä, mistä tietoa ei saa otettua strukturoidussa muodossa ja Big data-analyysissa keskeinen tehtävä onkin käsitellä strukturoimatonta dataa ja yrittää saada siitä analysoitavaksi hyödyllistä informaatiota.

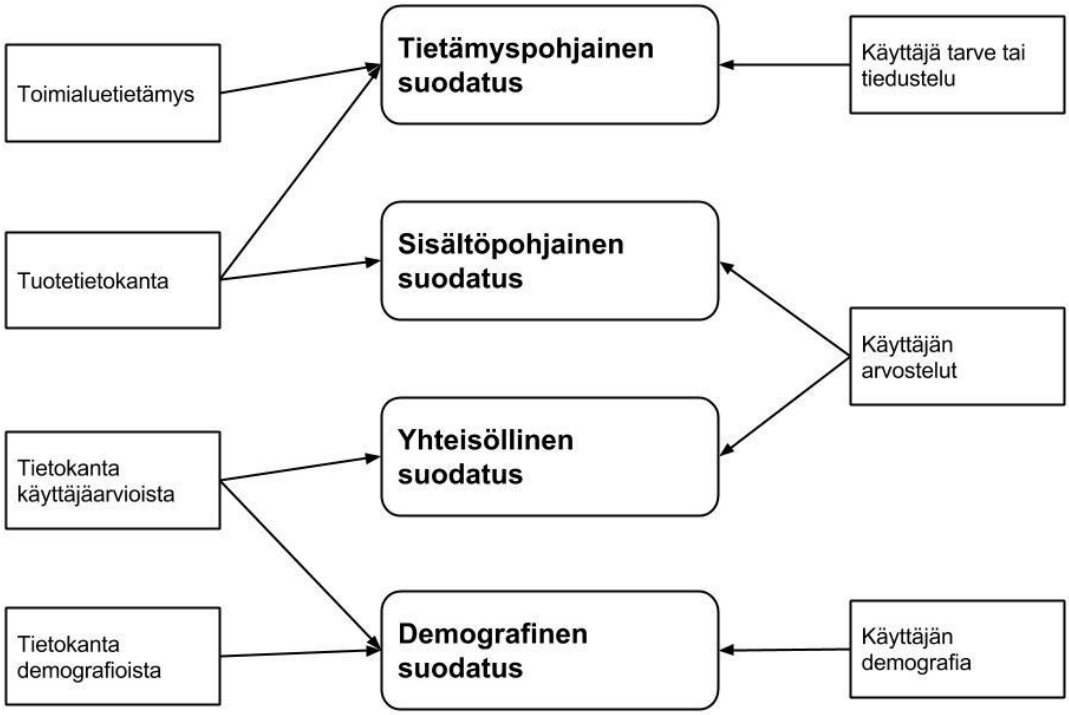
Miten Big data liittyy verkkokauppojen suosittelujärjestelmiin? Kuten luvussa 3 tullaan selittämään, suosittelujärjestelmät keräävät tietoa lukuisilla eri tavoilla ja monesta eri tietolähteestä. Osa tiedoista, kuten henkilötiedot ja muu demografinen profiili, on perinteisen datan tavoin hyvin jäsenneltyä. Käyttäjän selaustavat, selausaika ja selaushistoria verkkokaupan sivustoilla toimivat esimerkkinä implisiittisestä ja strukturoimattomasta datasta. Big datalle ominaisesti tietoa pitää käsitellä ennen kuin sitä voidaan analysoida suositteluja varten.

3 Työkalut suosittelujärjestelmien taustalla

Jotta suosittelujärjestelmän keskeisin tehtävä, käyttäjälle mieluisimpien tuotteiden tunnistaminen ja suositteleminen, toteutuisi, on suosittelujärjestelmän osattava ennustaa suositeltavat tuotteet (Ricci ym., 2011). Tämän aikaansaamiseksi suosittelujärjestelmän on kyettävä analysoimaan tuotteiden hyödyllisyys käyttäjää kohtaan. Menetelmiä tämän tehtävän toimittamiseksi on lukuisia ja ne vaihtelevat laajuudeltaan suuresti. Yksinkertaisimmillaan suosittelujärjestelmä ehdottaa suosituimpia tuotteita. Tämä on usein lähtökohta uudelle käyttäjälle, jonka makutottumuksia ja muuta metadataa ei ole vielä pystytty keräämään ja analysoimaan. Luotetaan siihen, että käyttäjä todennäköisesti pitää siitä mistä suurin osa muista käyttäjistä pitää.

Suosittelujärjestelmät on jaettu erilailla eri kirjallisuudessa, eikä yksiselitteistä tapaa tähän ole. Kahtena suurimpana suosittelujärjestelmätyyppinä pidetään yhteisöllistä sekä sisältöpohjaista suodatusta. Tämän lisäksi on esitelty muun muassa demografinen ja tietämyspohjainen suosittelujärjestelmä (Burke, 2007). Näiden lisäksi on olemassa hybridejä suosittelujärjestelmiä, joissa on yhdistetty joitain osia kahdesta tai useammasta suosittelujärjestelmästä sekä lisätty muita toiminnollisuuksia tarpeeseen paremmin räätälöidyn suosittelujärjestelmän aikaansaamiseksi.

Tässä luvussa käsitellään yleisimmät suosittelujärjestelmätyypit ja niiden toimintaperiaate yleisellä tasolla. Syventyminen teknisempään toimintarakenteeseen algoritmeja myöten ei ole mielekästä tämän laajuisessa tutkielmassa. Suosittelujärjestelmät on jaettu viiteen kategoriaan: yhteistoimintapohjaisiin, sisältöpohjaisiin, demografisiin, tietämyspohjaisiin sekä hybrideihin (kuvio 2). Hybridit suosittelujärjestelmät eivät näy kuviossa, sillä ne ovat muiden suosittelujärjestelmien yhdistelmiä.



KUVIO 2 Suosittelujärjestelmät ja niiden tietolähteet (Burke, 2007)

3.1 Yhteisöllinen suodatus (engl. Collaborative filtering)

Yhteistoimintapohjaista suodatusta (CF) pidetään ensimmäisenä automatisoituna suosittelujärjestelmänä (Konstan & Riedl, 2012) ja se on suosituin suosittelujärjestelmä (Sarwar, Karypis, Konstan & Riedl., 2001). CF ennustaa käyttäjälle suosituksia perustuen käyttäjän aikaisempiin arvioihin tuotteista tai palveluista, verraten näitä tietoja keskenään muiden saman palvelun käyttäjien kesken (Herlocker, Konstan, Terveen, Riedl., 2004). Käyttäjän antamia arvioita verrataan vertaiskäyttäjiin, jolloin CF ennustaa, voisiko käyttäjä pitää jostain hänelle uudesta tuotteesta (liite 1). Ennustaminen perustuu ihmisten toimintatapoihin ja käyttäytymiseen, ei koneen luomiin analyyseihin sisällöstä. CF:n lähestymistapa perustuu ns. ”joukon viisauteen” (engl. ”wisdom of the crowd”). CF:ää pidetään, erilaiset variaatiot mukaan lukien, yleisimpänä ja merkittävimpänä suosittelujärjestelmänä erityisesti kaupallisissa verkkopalveluissa. CF:n etuina nähdään sen mukautuvuus suodattaa sisältöä sisältötyypistä riippumatta, jonka seurauksena sitä pystytään soveltamaan useissa erilaisissa palvelukonsepteissa. CF kykenee tarjoamaan yllätyksellistä sisältöä sekä suodattamaan tietoa haastavien käsitteiden, kuten maun ja laadun mukaan (Herlocker ym., 2004.).

Schafer, Frankowski, Herlocker ja Sen (2007) listaavat kuusi tehtävää mihin CF soveltuu:

1. auttaa löytämään uusia tuotteita, mistä käyttäjä voisi pitää
2. neuvoo tietyn tuotteen kohdalla
3. auttaa löytämään toisen käyttäjän, kenellä on vastaavat intressit
4. auttaa ryhmää löytämään jotain uutta
5. auttaa löytämään sekoituksen tuotteita, jotka ovat vanhaa ja uutta
6. auttaa löytämään tuotteen/palvelun erityistä tarkoitusta varten.

CF toimintaperiaate voidaan jaotella kahteen pääkategoriaan; käyttäjäpohjaisiin (engl. user-based) ja tuotepohjaisiin (engl. item-based), koska nämä kaksi ovat suosituimmat CF kategoriat (Jian & Qun, 2012). Käyttäjäpohjaisessa menetelmässä pyritään etsimään käyttäjien välistä samankaltaisuutta, tuotepohjaisissa menetelmissä puolestaan tuotteiden välistä samankaltaisuutta (Jian & Qun, 2012). Toisaalta CF voidaan jakaa muistipohjaisiin (engl. memory-based) ja mallipohjaisiin (engl. model-based) suosittelujärjestelmiin. Muistipohjainen CF sisältää käyttäjäpohjaisen ja tuotepohjaisen CF:n ja tätä hyödyntävät esimerkiksi suuret verkkokaupat Amazon sekä Barnes & Noble omilla variaatioillaan (Su & Khoshgoftaar, 2009) Yhteisöllisen suodatuksen yhtenä suurimpana ongelmana nähdään haasteet suositella uusia tuotteita ja mitä suositella palvelun uusille asiakkaille. Tätä ongelmaa kutsutaan kylmäkäynnistysongelmaksi (Givon, 2011).

3.1.1 Muistipohjainen CF

Sekä käyttäjäpohjainen että tuotepohjainen menetelmä kuuluvat muistipohjaiseen CF:ään ja perustuu alun perin lähin naapuri-algoritmiin (engl. nearest-neighbor). Lähin naapuri-algoritmi luottaa ajatukseen, että käyttäjällä on samat makumieltymykset niin nykyhetkellä kuin tulevaisuudessakin, jolloin tuotteiden soveltuvuutta käyttäjälle voidaan ennustaa vertaamalla muiden samanlaisten käyttäjien makutottumuksia. Käyttäjien samankaltaisuutta voidaan estimoida esimerkiksi Pearsonin korrelaatiolla, joka laskee käyttäjän ja sen naapurikäyttäjien välistä korrelaatiota koskien tuotteiden arvottamista. Käyttäjäpohjainen lähin naapuri-algoritmiin perustuva yhteisöllinen suosittelumalli on yksi yleisimmistä variaatioista ja se on ensimmäinen CF toimintatapa yhdessä tuotepohjaisen CF:n kanssa. Käyttäjäpohjainen CF luottaa oletukseen että mikäli kaksi henkilöä ovat arvioineet usean tuotteen vastaavalla tavalla, tulevat he arvioimaan myös muita tuotteita vastaavasti. Tätä voidaan havainnollistaa yksinkertaisella esimerkillä:

Alla olevassa taulukossa (taulukko 1) käyttäjä Alice on arvioinut neljä tuotetta. Näiden arvioiden sekä neljän muun käyttäjän kaikille viidelle tuotteelle antamien arvioiden perusteella voimme laskea Pearsonin korrelaatiokertoimen avulla, pitääkö Alice Tuote 5:stä ja minkä arvosanan hän todennäköisesti sille antaa. Pearsonin korrelaatiolaskelmassa tulos on muuttujien kovarianssi välille [-1,1], arvon 1 ilmaistessa muuttujien välistä täydellistä riippuvuutta.

a,b : käyttäjät

$r_{a,p}$: käyttäjän a tuotteelle p antama arvo
P : joukko tuotteita, jotka on arvostellut sekä a että b

$$sim(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P}(r_{a,p} - \bar{r}_a)^2}\sqrt{\sum_{p \in P}(r_{b,p} - \bar{r}_b)^2}}$$

KUVIO 3 Pearson korrelaatio

TAULUKKO 1 Käyttäjäpohjainen CF.

	Tuote 1	Tuote 2	Tuote 3	Tuote 4	Tuote 5	Kovarianssi
Alice	5	3	4	4	?	
Käyttäjä 1	3	1	2	3	3	0,85
Käyttäjä 2	4	3	4	3	5	0,71
Käyttäjä 3	3	3	1	5	4	0
Käyttäjä 4	1	5	5	2	1	– 0,79

Laskemalla käyttäjien väliset korrelaatiot tuotteille 1-4, voimme nähdä että Alicen mieltymykset vastaavat parhaiten käyttäjää 1 (isoin kovarianssi), jolloin todennäköisin Alicen antama arvosana tuotteelle 5 on 3. Yllä oleva taulukko ei todellisuudessa olisi vielä tarpeeksi kattava, vaan lähin naapuri-algoritmissa 20–50 henkilön otanta nähdään riittäväksi tarkan ennusteen kehittämiseksi (Herlocker, Konstan & Riedl., 2000).

Muistipohjaisen CF:n ongelmana nähdään sen huono skaalautuvuus isoihin järjestelmiin, sillä suurien taulukoiden ylläpitäminen ei ole tehokasta palvelun kasvaessa riittävän suureksi. Muistipohjainen CF ennustaa tuotteita koko käyttäjä-tuote taulukosta (Sarwar ym., 2001). Esimerkiksi suurilla verkkokaupoilla voi olla miljoonia käyttäjiä ja tuotteita. Suurissa palveluissa naapuripohjainen suosittelujärjestelmä osaa suositella pääosin suosituimpia tuotteita jättäen vähemmän myydyt tuotteet piiloon suosituksilta niihin kertyneen vähäisen datan johdosta (Sarwar ym., 2001). Tällaisia tapauksia varten CF:stä on luotu lukuisia erilaisia, tiettyyn käyttötarkoitukseen paremmin soveltuvia variaatioita. CF-paradigmaa suosittelujärjestelmässään käyttävistä kaupallisista palveluista suurimpana pidetään verkkokauppaa Amazon.com, jota varten on kehitetty huomattavan suuren käyttäjäkunnan takia tarpeisiin paremmin soveltuva tuote-tuote (engl. item-item) CF. (Konstan ym., 2012). Monissa verkkokaupoissa tuotteita on vähemmän mitä käyttäjiä, jolloin tuote-tuote CF toimii perinteistä käyttäjä-käyttäjä CF:ää nopeammin.

Toinen suuri ongelma muistipohjaisessa CF:ssä on kylmäkäynnistysongelma (engl. cold-start problem, data sparsity problem) (Su & Khoshgoftaar, 2009). Muistipohjaisen CF:n perustuessa muiden käyttäjien antamiin arvioihin, esimerkiksi verkkokauppaan lisättyä uutta tuotetta ei voida suositella yhdellekään käyttäjälle, sillä se ei ole saanut yhtäkään arviota. Ongelma koskee myös uutta käyttäjää. Palvelun uusi käyttäjä ei ole ehtinyt arvostella yhtään tuotetta,

jolloin käyttäjäprofiilia ei ole ehtinyt muodostumaan, mikä tekee käyttäjän mieltymysten vertailun muihin käyttäjiin mahdottomaksi. Toisin sanoen naapurustoa kyseiselle käyttäjälle ei ole (Schafer ym., 2007). Tähän varsin suureen ongelmaan on olemassa useita parannusvaihtoehtoja, esimerkiksi mallipohjainen CF sekä muut hybridit ratkaisut unohtamatta toisenlaisia suosittelujärjestelmiä kuten sisältöpohjaiset suosittelujärjestelmät. Käyttäjää voidaan myös ”pakottaa” arvioimaan joitain tuotteita ensimmäisen käyttökerran yhteydessä käyttäjäprofiilin luomiseksi. Yksi vaihtoehto uusien tuotteiden kohdalla on antaa niille jokin oletusarvo, jolloin tuotteet saavat paremmin näkyvyyttä.

3.1.2 Mallipohjainen CF

Muistipohjaisen CF:n heikkouksia, erityisesti kylmäkäynnistysongelmaa, on pyritty eliminoimaan luomalla yhteisöllisestä suodatuksesta mallipohjainen (engl. model-based) versio. Mallipohjaisen CF:n toimintaperiaate on tarjota tuotesuositteluja luomalla aluksi malli käyttäjän antamista arvioista. Mallipohjaisen CF:n algoritmit ottavat ennustavan lähestymistavan ja visioivat yhteistoiminnallisen prosessin laskemalla oletusarvot tuotteille. Mallipohjainen CF nimensä mukaisesti hyödyntää muiden käyttäjien ja tuotteiden arvioita ennustuksissaan, mutta se osaa, suosittelujärjestelmätyypistä riippuen, ennustaa ja laskea etukäteen tiettyjä oletusarvoja käyttäen koneoppimis- algoritmeja (engl. machine learning), esimerkiksi Bayesin luokittelijaa (engl. Bayesian network), klusterointia tai sääntöpohjaista lähestymistä (Sarwar ym., 2001).

Bayesin luokittelijan toimintaperiaate perustuu todennäköisyyksien laskemiseen (Paaso, 2012). Ideana on luoda päätöspuu palvelun tuotteista. Jos käyttäjä on katsonut elokuvan x ja y ja hänen ”naapurustossa” elokuvat x ja y nähneet ihmiset ovat myös katsoneet elokuvan z, käyttäjä todennäköisesti haluaa katsoa elokuvan z. Tällöin elokuvaa z voidaan suositella käyttäjälle (Schafer ym., 2007). Bayesin luokittelija on tutkittu toimivan tehokkaaksi tuotteiden suosittelussa, mutta se ei tuo lisäarvoa suosituksissa, jossa käsitellään useita arvoja kuten elokuvan arvosanaa (Schafer ym., 2007). Joissain Bayesin luokittelijan variaatioissa moniarvoiset taulukot on muutettu binäärimuotoon ennustuksen toteuttamiseksi (Su & Khoshgoftaar, 2009). Tämä tekniikka on osoittautunut muistipohjaista CF:ää paremmin skaalautuvaksi, mutta ennustustarkkuus on heikompi. Lisäksi Bayesin luokittelijan haittapuolena on laskennallisesti erittäin kallis mallin opetus ja päivitys, mikäli käyttäjiä on paljon (Paaso, 2012).

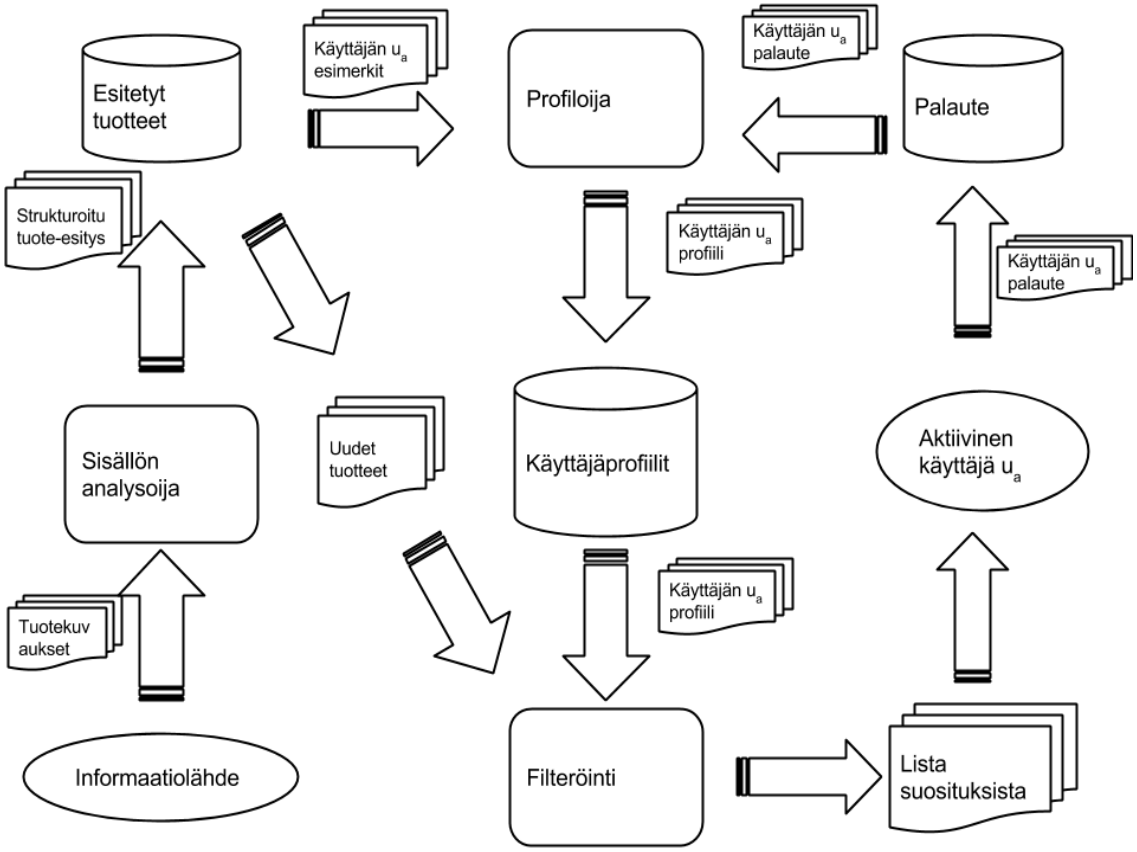
Klusteroinnissa perusajatus on yhdistää samankaltaiset dataobjektit yhteen klusteriin. Klusterointia pidetään monesti vain eräänlaisena välivaiheena ja lopullista analyysia varten käytetään jotain muistipohjaista menetelmää, kuten Pearsonin korrelaatiota (Su & Khoshgoftaar, 2009). Klusteroinnin etuna on perinteistä CF:ää huomattavasti parempi skaalautuvuus, sillä ennustukset tehdään pienissä osissa eikä kokonaisessa käyttäjä-tuote-tietokannassa. Ennustuksen laatu on kuitenkin yleisesti melko heikko.

3.2 Sisältöpohjainen suodatus (engl. Content-based filtering)

Sisältöpohjaisen suodatuksen (CBF) perusoletus on, että tuotteet vastaavilla ominaisuuksilla arvostellaan samankaltaisesti (Schafer ym., 2007). Tuotteille on merkitty tiettyjä ominaisuuksia ja CBF vertailee näitä käyttäjien tekemiin valintoihin osatakseen suositella vastaavia tuotteita (Ricci ym., 2011). Jos käyttäjä on katsonut komedia-genreen kuuluvaa tv-sarjaa, hän saattaa tykätä muista vastaavan kaltaisista tv-sarjoista (liite 2). CBF ei täten tarvitse arvosteluja tuottaakseen suosituksia käyttäjälle. CBF tekee suosituksia analysoimalla sisältöä esimerkiksi asiakirjoista, URL-osoitteista, uutisviesteistä, tuotteen kuvauksesta tai tiedoista käyttäjän profiilissa (Su & Khoshgoftaar, 2009). Näiden tietojen avulla CBF luo käyttäjäprofiilin. Käyttäjästä tehdyn profiilin attribuutteja verrataan myytävien tuotteiden attribuutteihin, jolloin löydetään käyttäjää todennäköisesti kiinnostavat tuotteet (Ricci ym., 2011). Suosittelut voivat olla erittäin osuvia, mikäli käyttäjäprofiili vastaa hyvin käyttäjän oikeita intressejä.

Käyttäjä on voinut ostaa tuotteita, mitkä eivät lopulta olleetkaan hänen mielestään mielekkäitä (esimerkiksi katsonut ohjelman Netflixistä, mikä ei ollut mielenkiintoinen). Nämä hankinnat kuitenkin vaikuttavat profiilin luomisessa, jolloin CBF tarjoaa käyttäjälle vastaavanlaisia, ei mielenkiintoisia tuotteita. Tämän kaltainen ongelma voidaan kuitenkin ehkäistä pyytämällä käyttäjää arvioimaan ostamansa tuotteet, jolloin CBF osaa antaa attribuuteille painoarvon. Tutkittavia elementtejä analysoidaan niiden tärkeyden perusteella tietyssä kontekstissa (esim. kuinka usein jokin sana toistuu internet-sivustolla) ja verrataan aiemmin käyttäjästä kerättyyn informaatioon.

CBF:n toimintaperiaate koostuu kolmesta pääkomponentista (kuvio 3): Sisällön analysoijasta, profiloijasta sekä filteröinnistä. Prosessi alkaa sisällön analysoijasta, minkä tehtävänä on analysoida uutta, strukturoimatonta sisältöä (esim. dokumentit, tuotekuvaukset) muotoon, jota voidaan hyödyntää seuraavissa vaiheissa. Tämä strukturoitu tieto lähetetään eteenpäin profiloija- ja filteröinti-komponenteille. Profiloija kerää dataa käyttäjän toiminnasta ja makumieltymyksistä ja yrittää yleistää tätä tietoa käyttäjäprofiilin muodostamiseksi. Käyttäjästä kerätyn tiedon yleistämisessä käytetään usein koneoppimisalgoritmeja. Filteröinti-komponentti hyödyntää käyttäjäprofiilia vertaillakseen sitä sisällön analysoijalta saamiinsa uusiin tuotteisiin. Käyttäjän saadessa suosituksia filteröinti-komponentilta, hän voi antaa siitä halutessaan palautetta. Palaute (esim. arvosana tuotteesta) siirtyy profiloija-komponentin käsiteltäväksi tehden käyttäjäprofiilista tarkemman.



KUVIO 4 Sisältöpohjaisen suodatuksen arkkitehtuuri (Ricci ym., 2011)

Palautetta voidaan kerätä käyttäjältä kahdella tapaa, eksplisiittisesti ja implisiittisesti. Eksplisiittisessä tavassa tietoa pyydetään käyttäjältä jotain tuotetta koskien. Tätä tapaa hyödynnetään esimerkiksi tilanteissa, missä uusi käyttäjä on kirjautunut palveluun eikä hänestä ole vielä tarvittavaa määrää tietoa suosittelujen tuottamiseksi. Implisiittisessä palautteen keräämisessä seurataan käyttäjän toimintaa kuten tietyn tuotteen katsomiskertoja, tuotteen tallentamista kirjanmerkkeihin jne. Implisiittisessä tavassa käyttäjää ei siis pyydetä tekemään mitään, vaan tieto saadaan kerättyä seuraamalla käyttäjän toimintaa (Ricci ym.,2011).

Sisältöpohjaisen suodatuksen edut ja haitat

Vertailtaessa sisältöpohjaista suodatusta yhteistoimintapohjaiseen, Ricci ym. (2011) listasivat seuraavat edut:

- Riippumattomuus muista käyttäjistä: CBF toimii käyttäjän oman toiminnan ja tuotteille antamien arvioiden perusteella luoden käyttäjästä oman profiilin. Muiden ihmisten arviointeja tuotteista ei täten tarvita.
- Läpinäkyvyys: Käyttäjän luottamista palvelua kohtaan lisää toimintojen läpinäkyvyys. CBF voi näyttää käyttäjälle, miksi jotain tuotet-

ta hänelle suositellaan (”koska pidit tuotteesta x, voisit pitää myös tuotteesta y”). Tällaista ei voida toteuttaa yhteistoimintapohjaisessa suodatuksessa, sillä suosittelut perustuvat muiden, anonyymien käyttäjien antamiin arvioihin.

- Uudet tuotteet: CBF kykenee suosittelemaan tuotetta, jota ei ole vielä arvioitu kenenkään käyttäjän toimesta. Tällöin CBF ei kärsi niin pahasti kylmäkäynnistysongelmasta. Toimiakseen alusta alkaen CBF tarvitsee kuitenkin tietoa käyttäjästä.

CBF kärsii kuitenkin muutamista haitoista:

- Rajallinen sisällön analyysi: Saadakseen tuotteista tietoa, CBF pitää tietää kontekstiin liittyviä asioita. Esimerkiksi kerätessä tietoa elokuvista, CBF on tunnistettava näyttelijät ja ohjaajat. Luonnollisesti relevantin tiedon automaattisessa keräämisessä tulee raja vastaan. Manuaalinen tiedon keräys on puolestaan hidasta ja resurssimielessä tehotonta.
- Yli-erikoistuminen: CBF:llä on taipumusta muokkautua liian spesifiksi käyttäjän makutottumuksia kohtaan, sillä CBF tutkii käyttäjän aiempaa käyttäytymistä (Su & Khoshgaftaar, 2009). Jos käyttäjä on arvioinut vain tietynlaisia teoksia, CBF osaa suositella vain vastaavankaltaisia jättäen pois tuotteet muista kategorioista.
- Uusi käyttäjä: Myös CBF kärsii kylmäkäynnistysongelmasta koskien uutta käyttäjää. Uuden käyttäjän on arvioitava jotain tietoa, jotta CBF kykenee antamaan tuotesuositteluja.

3.3 Demografinen suosittelujärjestelmä

Nimensä mukaisesti demografisen suosittelujärjestelmän toimintaperiaate perustuu käyttäjän demografiseen profiiliin (Burke, 2007). Demografinen suosittelujärjestelmä kategorisoi käyttäjän hänen demografisten tietojen (esim. ikä, sukupuoli, kansalaisuus) perusteella ja suosittelee tuotteita demografisten luokkien mukaan (Burke, 2002). Tätä suosittelujärjestelmää käytetään useilla www-sivustoilla jollain tasolla. Esimerkkinä maan ja kielen valinta ennen varsinaiselle sivustolle pääsemistä suodattaa käyttäjän hänen kansalaisuudelleen räätälöidylle sivulle (Ricci ym., 2011). Demografisen suosittelujärjestelmän etuna nähdään, että se osaa ehdottaa käyttäjälle tuotteita alusta alkaen, tosin käyttäjästä on oltava demografisia tietoja kerättynä (Burke, 2002).

3.4 Tietämyspohjainen suosittelujärjestelmä (engl. Knowledge-based)

Tietämyspohjainen suosittelujärjestelmä ehdottaa tuotteita käyttäjän tarpeista ja mieltymyksistä päätellen (Burke, 2007). Suosittelu perustuu erityiseen toimialuetietämykseen (engl. domain knowledge) kuinka tuotteen ominaisuudet sopivat yhteen käyttäjän tarpeiden ja mieltymysten kanssa (Ricci ym., 2011). Esimerkkinä voidaan pitää palveluita, jotka etsivät käyttäjälle jotain tuotetta vastaavat tuotteet (liite 3). Esimerkiksi käyttäjä voi hakea yhden elokuvan perusteella sitä vastaavat elokuvat. Mikäli haku ei miellytä tarpeeksi, voidaan siihen lisätä jokin attribuutti, joka ohjaa hakua haluttuun suuntaan (esim. painotus tiettyä lajityyppiä kohtaan). Tätä toimintoa kutsutaan samankaltaisuushauksi (Burke, 2000). Tietämyspohjaiset suosittelujärjestelmät ovat usein tapauspohjaisia (case-based) tai rajoitepohjaisia (constraint-based). Molemmissa tiedonkeruu käyttäjältä toimii vastaavanlaisesti, mutta tiedonkäsittelymetodeissa on eroavaisuuksia.

Tietämyspohjaiset suosittelujärjestelmät toimivat hyvin alusta alkaen ja sen suurimpana etuna pidetäänkin kylmäkäynnistysongelman puuttuminen. Toimintaperiaate ei vaadi käyttäjän tai naapuruston arvioita tuotteista. Toimiakseen tehokkaasti myös jatkossa, suosittelujärjestelmässä on oltava tehokkaita oppimiskomponentteja (Ricci ym., 2011).

3.5 Hybridit suosittelujärjestelmät

Suurin motivaatio hybridien suosittelujärjestelmien kehittämiseen on aiemmin kehitetyn suosittelujärjestelmän puutteiden paikkaaminen. Tämä toteutetaan yhdistelmällä osia kahdesta tai useammasta suosittelujärjestelmästä (Ricci ym. 2011). Esimerkiksi kylmäkäynnistysongelmista pahasti kärsivää yhteistoimintapohjaista suosittelujärjestelmää voidaan parantaa ottamalla ominaisuuksia sisältöpohjaisesta suosittelujärjestelmästä, missä suosittelua tehdään tuotteiden ominaisuuksien perusteella. Hybridien suosittelujärjestelmien etuna ovat käyttötarkoitukseen tarkemmin sopivat ominaisuudet, toisin sanoen suosittelujärjestelmästä saadaan yhdistelemällä huomattavasti spesifimpi. Yhdistetyt suosittelujärjestelmät voivat myös olla toimintaperiaatteeltaan samoja. Esimerkiksi kaksi eri sisältöpohjaista suosittelujärjestelmää voidaan yhdistää yhdeksi hybridiksi järjestelmäksi (Burke, 2007). Toinen voi käyttää data-analyysissa Bayesin luokittelijaa, toinen lähin naapuri-algoritmia.

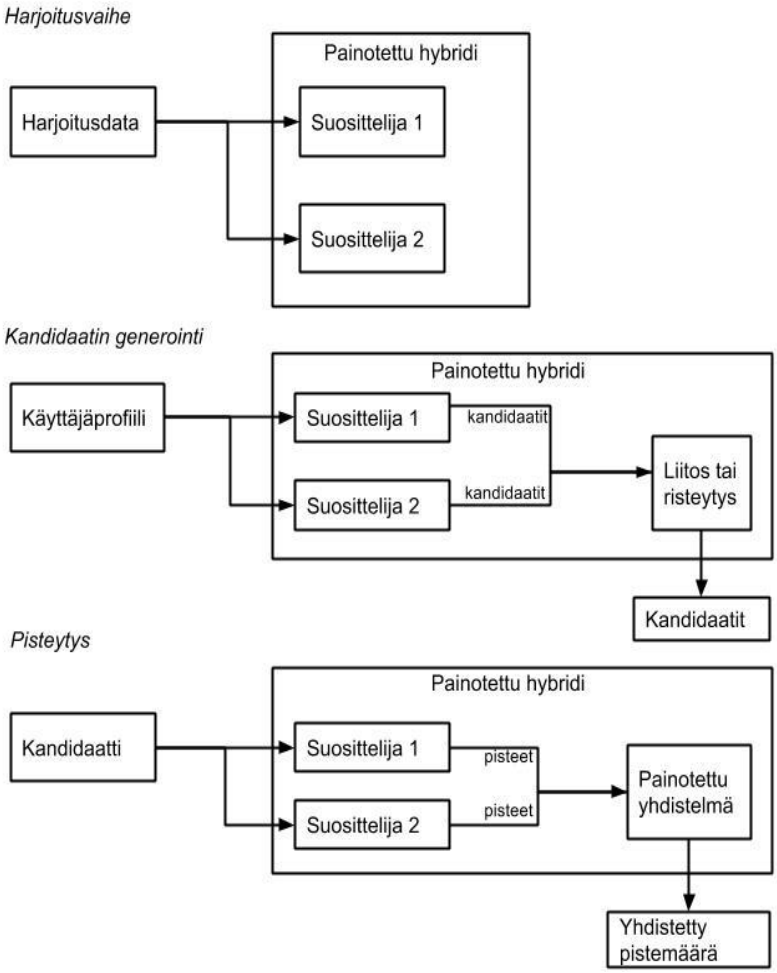
Hyvänä käytännön esimerkkinä hybridistä suosittelujärjestelmästä on videoiden suoratoistopalvelu Netflix. Yhteistoimintapohjaisen suosittelujärjestelmän tavoin se tekee suosituksia vertailemalla käyttäjän katsomishistoriaa vertaiskäyttäjiin. Tämän lisäksi suosituksia tehdään käyttäjän itse antamien ar-

vosanojen perusteella vertailemalla näitä samankaltaisiin artikkeleihin, sisältöpohjaisen suosittelujärjestelmän tavoin.

Burke (2007) lajittelee hybridit suosittelujärjestelmät seitsemään eri kategoriaan:

- painotettu (engl. weighted)
- kytkentä (engl. switching)
- sekoitettu (engl. mixed)
- ominaisuuksien yhdistäminen (engl. feature combination)
- ominaisuuksien lisääminen engl. (feature augmentation)
- sarja (engl. cascade)
- meta-taso (engl. meta-level).

Useimpien hybridien suosittelujärjestelmien toimintatapa koostuu kolmesta päävaiheesta: harjoitusvaiheesta, kandidaatin luomisesta ja pisteytysvaiheesta (Burke, 2007). Alla on esimerkkinä kaavio painotetun hybridin suosittelujärjestelmän toimintaperiaatteesta (kuvio 4).



KUVIO 5 Painotettu hybridi suosittelujärjestelmä (Burke, 2007)

Painotettu hybridi suosittelujärjestelmä edustaa hybridijärjestelmää yksinkertaisimmillaan. Harjoitusvaiheessa jokainen suosittelukomponentti käy lävitse harjoitusdataa. Tämä vaihe sisältyy lähes jokaiseen hybridin suosittelujärjestelmän seitsemästä kategoriasta. Toisessa vaiheessa suosittelukomponentit ehdottavat kandidaattituotteita. Tämän vaiheen toimintaperiaate riippuu käytettävistä suosittelujärjestelmistä. Esimerkiksi sisältöpohjainen suosittelujärjestelmä käy läpi käyttäjäprofiilia verraten tätä dataa tuotetietokantaan. Vertailu tuo esille kandidaattituotteita, jotka käsitellään erilailla riippuen käytettävästä toimintaperiaatteesta. Kandidaattituotteet ovat varteenotettavia vaihtoehtoja, ei vielä suositeltavia kohteita. Lopputulemana ovat kandidaattituotteet, jotka siirtyvät pisteytysvaiheeseen. Pisteytysvaiheessa jokainen hybridin suosittelujärjestelmän komponentti pisteyttää kandidaattituotteen. Nämä pisteytykset lasketaan yhteen jolloin saadaan yhdistetty pisteytys. Pisteytyksen jälkeen suositeltavat tuotteet lajitellaan pisteytyksen mukaan laskevassa järjestyksessä ja näytetään käyttäjälle. Painotuksella viitataan suosittelukomponenttien painoarvoon suosituksissa. Kaksi eri suositelutekniikkaa voidaan esimerkiksi painottaa suhteessa 60/40, jolloin ensimmäisen suosittelukomponentin tulokset merkitsevät jälkimmäistä enemmän (Burke, 2007).

Kytkentä-kategoriaan kuuluvat hybridit suosittelujärjestelmät analysoivat käyttäjäprofiilin ja valitsevat vain yhden käyttötarkoitukseen parhaiten soveltuvan suosittelujärjestelmän. Perusajatus on, että tiettyyn tilanteeseen toinen suosittelujärjestelmä sopii paremmin kuin toinen. Teknisesti haastava vaihe on tapa, jolla oikea suosittelujärjestelmä valitaan.

Sekoitus-vaihtoehdossa kaksi tai useampi suosittelukomponenttia luo listan kandidaateista, jotka pisteytetään saman suosittelukomponentin toimesta. Kandidaatteja ei näin ollen yhdistetä missään vaiheessa, vaan ne etenevät rinnakkain. Pisteytysvaiheen lopuksi järjestetty lista suosituksista yhdistetään yhdeksi näkymäksi. Tämän haasteena on esitystapa. Missä järjestyksessä tuotteet näytetään, miten käsitellä mahdolliset duplikaatit jne.

Hybrideistä suosittelujärjestelmistä ominaisuuksien yhdistäminen-menetelmä eroaa aiemmin esitetyistä melko paljon. Varsinainen suosittelu tapahtuu vain yhden suosittelukomponentin voimin. Hybridin tästä järjestelmästä tekee se, että alkuvaiheessa käsiteltävää dataa analysoidaan algoritmein, johon on lisätty toiminnollisuuksia toisesta suosittelujärjestelmämenetelmästä. Jos varsinainen suosittelu tapahtuu sisältöpohjaisesti, on datan analysointiin alkuvaiheessa voitu ottaa elementtejä yhteistoimintapohjaisesta suosittelujärjestelmästä.

Ominaisuuksien lisääminen-menetelmä muistuttaa osin ominaisuuksien yhdistämistä. Jokaisessa kolmessa vaiheessa käsiteltävä data siirtyy aluksi avustavalle suosittelijalle, joka esikäsittelee dataa ja tarvittaessa lisää siihen sisältöä. Tämä ”kasvatettu” data välittyy varsinaiselle suosittelukomponentille.

Sarja-hybridi auttaa tilanteessa, missä perinteinen suosittelujärjestelmä antaa tuotteille samat pisteytykset. Tällaiset tilanteet nähdään ongelmallisina. Sarja-hybridi koostuu kahdesta suosittelukomponentista, päätoimisesta ja avustavasta. Avustava suosittelukomponentti puuttuu suositteluprosessiin vain, mi-

käli päätoiminen suosittelukomponentti antaa kahdelle tai useammalle tuotteelle samat pisteytykset.

Meta-taso-hybridijärjestelmä käyttää kahdessa ensimmäisessä, harjoittelu- ja kandidaatin generointivaiheessa, avustavaa suosittelukomponenttia ennen päätoimista suosittelukomponenttia. Erona ominaisuuksien lisääminen-hybridiin on se, että meta-taso-menetelmässä avustava suosittelukomponentti korvaa alkuperäisen datan täysin ja näin varsinainen suosittelukomponentti ei käsittele alkuperäistä dataa missään vaiheessa.

4 Suosittelujärjestelmien tarjoamia hyötyjä palveluntarjoajan näkökulmasta

> Vuonna 1988 Brittiläinen kiipeilijä Joe Simpson kirjoitti kirjan nimeltä Touching the Void, riipaiseva kertomus kuolemanläheisistä kokemuksista Perun Andeilla. Vaikka kirja-arvioinnit olivat hyviä, teos menestyi vain kohtalaisesti ja lukijat unohtivat teoksen nopeasti. Noin kymmenen vuotta myöhemmin sattui outo tapaus. Jon Krakauerin kirjoittama teos Jäätäviin korkeuksiin, kirja mikä myös kertoi vuorikiipeilystä, muodostui menestysteokseksi heti julkaisuhetkellä. Yhtäkkiä myös Touching the Void alkoi myydä uudestaan.
>
> Mitä tapahtui? Kuulopuheita verkossa. Muutamat Jäätäviin korkeuksiin-kirjan lukeneet olivat arvioineet sen Amazon.com-verkkokaupassa ja samalla viitanneet siinä olevan yhtäläisyyksiä Touching the Void-teoksen kanssa. Muut lukijat lukivat arvostelut ja lisäsivät ostoskoriinsa myös tämän teoksen. Amazon.comin suosittelujärjestelmä huomasi tämän yhtäläisyyden ja alkoi suosittelemaan kirjaa. ”Lukijat, jotka ostivat kirjan Jäätäviin korkeuksiin, ostivat myös teoksen Touching the Void.” Asiakkaat hyväksyivät ehdotuksen, ostivat teoksen ja kirjoittivat lisää positiivisia arvosteluja.
>
> Tämän tapauksen johdosta kirjaa Touching the Void on myyty enemmän kuin teosta Jäätäviin korkeuksiin. (Anderson, 2006, 25–26)

Verkkokauppojen tuotevalikoima mitataan usein miljoonissa kappaleissa. Oikean tuotteen löytäminen selailemalla ja haku-toimintoja käyttäen voi olla asiakkaan näkökulmasta turhauttavaa. Suosittelujärjestelmät syntyivät tähän tarpeeseen ja niiden suosio verkkokaupoissa on kasvanut räjähdysmäisesti vuosien saatossa (Schafer, Konstan & Riedl., 2001). Nykypäivänä on hankala löytää verkkokauppaa, missä ei olisi jonkin tasoista suosittelujärjestelmää. Suosittelujärjestelmien rooli verkkokaupassa on olla ”virtuaalinen myyjä”, joka osaa tarjota asiakkaalle parhaimmillaan erittäin yksilöllistä palvelua (Schafer, Konstan & Riedl., 1999) Suosittelujärjestelmien etuina on palvelun personointi asiakkaan mieltymysten mukaisesti, mikä omalta osaltaan auttaa luomaan lojaaleja asiakassuhteita. Voidaan ajatella, että suosittelujärjestelmät luovat kokonaan uuden

kaupan, mikä on räätälöity tietyn asiakkaan tarpeiden mukaiseksi (Schafer ym., 2001). Toinen merkittävä hyöty on sen tuoma tuotteiden näkyvyyden lisäys, myös vähemmän suosituille tuotteille. Tätä kutsutaan pitkäksi hännäksi, mistä tarkemmin luvussa 4.2.

Ricci ym. (2011, 5) listasivat seuraavia syitä suosittelujärjestelmien käyttöönottoon:

- lisää myytyjen tuotteiden määrää
- myy enemmän erilaisia tuotteita
- lisää asiakastyytyväisyyttä
- lisää asiakkaan lojaaliutta
- parempi ymmärrys asiakkaan tarpeista.

Schafer ym. (1999) mainitsevat kolme tapaa, miten suosittelujärjestelmät parantavat sähköisen liiketoiminnan myyntiä:

- Selailijoista ostajia: Verkkosivuilla vierailee usein käyttäjiä, jotka eivät ole ostoaikeissa. Suosittelujärjestelmät tuovat paremmin esille tuotteita, joita käyttäjä päätyy lopulta ostamaan.
- Ristiinmyynti: Suosittelujärjestelmät parantavat ristiinmyyntiä ehdottamalla tuotteita, jotka liittyvät jo ostoskoriin lisättyihin tuotteisiin. Esimerkiksi käyttäjä, joka on ostamassa dvd-soitinta, voi olla halukas ostamaan myös HDMI-kaapelin.
- Lojaalius: Lojaaliuden saaminen on verkkoliiketoiminnassa erittäin tärkeää, sillä kilpailevat palvelut ovat vain muutaman klikkauksen päässä. Sivustolla kävijän saaminen uskolliseksi asiakkaaksi helpottuu suosittelujärjestelmän oppiessa käyttäjästä tarpeellisia tietoja henkilökohtaisen palvelun tarjoamiseksi. Mitä paremmin verkkokauppa osaa suositella tuotteita asiakkaan todellisia mieltymyksiä vastaan, sitä todennäköisemmin hän palaa takaisin.

4.1 Suosittelujärjestelmät ja pitkä häntä

Suosittelujärjestelmien taloudellisia etuja tutkiessa ei voi olla törmäämättä ilmiöön nimeltä pitkä häntä. Pitkä häntä (engl. The long tail) on vuonna 2004 Wired-lehden päätoimittajan Chris Andersonin laatima termi määrittämään nichetuotteiden myynnin kasvua erityisesti verkkokauppojen ja suoratoistopalveluiden ansiosta. Anderson on kirjoittanut ilmiöstä kirjan nimeltä Pitkä häntä: Miksi tulevaisuudessa myydään vähemmän enempää (2006).

Anderson tutki vuonna 2004 suurten verkkoyritysten datan käyttöä. Luonnollisesti yritykset eivät julkaise keräämänsä datan sisältöä ja tästä syystä Anderson haastatteli yritysten toimitusjohtajia. Hän aloitti työnsä haastattelemalla Ecast-nimistä digitaalista jukebox-palvelua tarjoavan yrityksen toimitusjohtajaa Robbie Vann-Adibéa. Haastattelussa kävi ilmi, että yrityksen noin

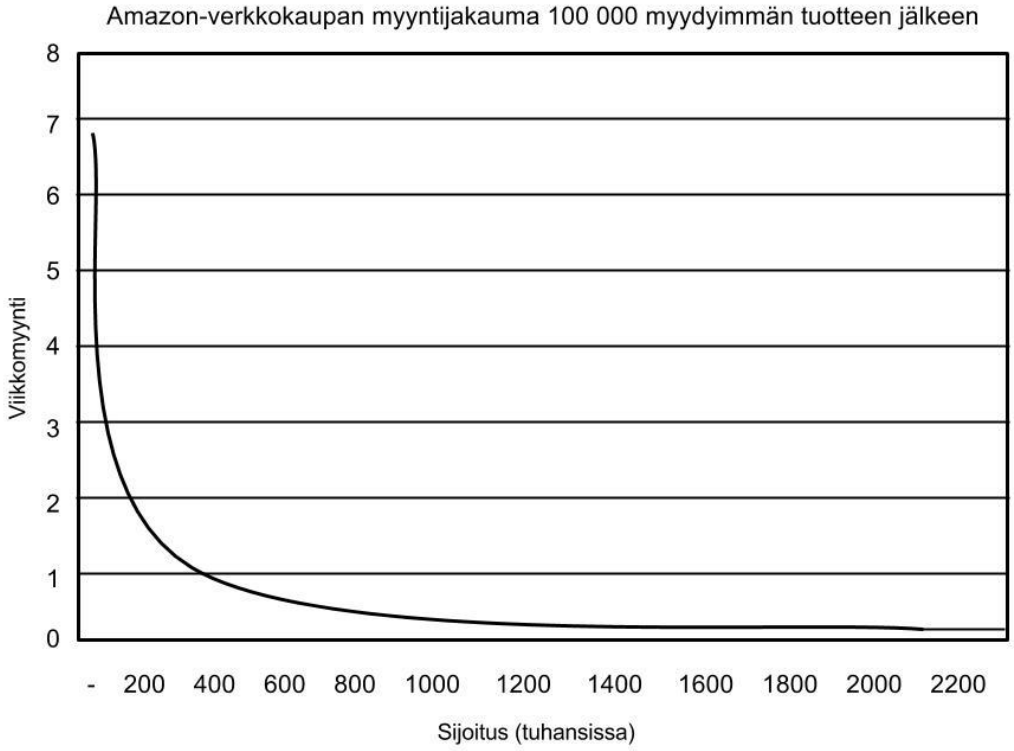
10 000 musiikkikappaleesta 98% on sellaisia, joita ostetaan vähintään kerran kvartaalissa. Luku hämmästytti Andersonia, joka oli arvellut luvun huomattavasti pienemmäksi ja rikkoi nk. Pareton periaatetta (Anderson, 2006, 16–20). Vilfredo Pareton 80/20-sääntöä sovellettuna liiketoimintaan, 20 % suosituimpia tuotteita muodostavat 80 % myynnistä. Tästä asiasta innostuneena Anderson jatkoi tutkimuksiaan haastattelemalla muun muassa Amazonia, iTunesia ja Netflixiä. Kaikissa mainituissa palveluissa oli vastaava ilmiö: noin 98 % tuotteista ovat sellaisia, joita myydään vähintään kerran kvartaalissa. Tätä kutsutaan 98 % säännöksi (engl. 98 percent rule).

Pitkä häntä ilmenee kaupoissa ja palveluissa, joissa on valtava tuotevalikoima. Käytännössä pitkä häntä-ilmiötä ilmenee verkkokaupoissa. Suuri varastotila kiinteistökustannuksiltaan edullisella alueella, pienet henkilöstökustannukset, tehokas logistiikka muiden perinteistä kivijalkakauppaa alempien kustannusten ohella mahdollistavat lähes rajattoman määrän hyllytilaa. Tällöin kaupan on mahdollista ottaa valikoimiinsa huomattavasti enemmän tuotteita, myös sellaisia mitä perinteisellä kaupalla ei ole varaa ottaa. Brynjolfsson, Hu ja Smith (2003) käyttävät esimerkkinä Amazon.com verkkokauppaa verrattaessa tuotevalikoimaa perinteiseen kivijalkakauppaan (taulukko 2).

TAULUKKO 2 Verkkokaupan ja kivijalkamyymälän ero tuotevalikoimassa (Brynjolfsson ym., 2003)

Tuotekategoria	Amazon.com	Perinteinen kivijalka-myymälä
Kirjat	2 300 000	40 000 – 100 000
CD:t	250 000	5 000 – 15 000
DVD:t	18 000	500 – 1 500
Digitaalikamerat	213	36
MP3-soittimet	128	16
Skannerit	171	13

Alla näemme Amazon-verkkokaupan myyntijakauman myydyimmän 100 000 tuotteen jälkeen (kuvio 5). Kuvasta näemme, että myynti luonnollisesti laskee dramaattisesti suosituimpien tuotteiden jälkeen, mutta huomattavaa on, että myynti ei koskaan lopu. Termi saa nimensä juuri tämän kaltaisesta kuvaajasta, missä tarjonnan häntä on muuttunut pitkäksi. Pareton periaate ei enää toteudukaan, vaan suuri osa myynnistä (Amazonin tapauksessa 30–40 %) muodostuu suuresta määrästä tuotteita, joita yksinään myydään vähän.



KUVIO 6 Pitkä häntä Amazon-verkkokaupassa (Brynjolfsson ym., 2003)

Anderson (2006, 258) on jakanut neljän suuren internetissä toimivan yrityksen myyntijakauman kahteen tuoteosioon: 100 myydyintä ja 101:stä eteenpäin ilmoittaen näiden prosenttiosuuden kokonaismyynnistä (taulukko 3). Huomionarvoista taulukossa on sadan myydyimmän tuotteen ulkopuolelle jäävien tuotteiden myyntiosuus kokonaismyynnistä. Esimerkiksi Netflixin kohdalla myynnistä suurempi osa muodostuu näistä tuotteista.

TAULUKKO 3 Myyntijakauma (Anderson, 2006, 258)

Sijoitus	Wal-Mart	Rhapsody	Blockbuster	Netflix
1-100	65 %	47 %	68 %	38 %
101 →	36 %	53 %	32 %	62 %

Brynjolfsson, Hu ja Smith (2006) jakavat pitkän hännän anatomian neljään osaan: tarjontapuoleen ja kysyntäpuoleen jakaen molemmat ensimmäiseen ja toiseen tasoon (taulukko 4). Tarjontapuoli näyttää, mitä vaaditaan pitkän hännän kaltaisen tarjontamallin saavuttamiseksi ja mitkä ovat sen hyödyt. Toinen taso kuvaa, mitä ilmiöitä ensimmäisen tason muutokset muodostavat.

Kysyntäpuoli ilmaisee, mitkä tekijät ajavat kuluttajia käyttäytymään siten, että pitkä häntä-tarjontamalli tapahtuu. Se sisältää kuluttajan kannalta aktiivisia

ja passiivisia tekijöitä kuten hakutyökalujen käyttämistä ja suosittelujärjestelmiä. Toinen taso kuvastaa muutoksia johtuen ensimmäisen tason tekijöistä.

TAULUKKO 4 Pitkä häntä-käsitteen anatomia (Brynjolfsson ym., 2006)

	1. taso	2. taso
Tarjontapuoli (tuottajat, jälleenmyyjät)	• kustannus: virtuaalinen hyllytila, tilausohjautuva tuotanto, sähköinen toimitus jne. • hyödyt: kuluttajien yhdistäminen	• lisääntynyt kannuste uusien tuotteiden kehittämiseen • markkinointistrategioiden muokkaaminen pitkä häntä-mukaiseksi • uusia teollisuusrakenteita
Kysyntäpuoli (kuluttajat)	• aktiivinen: tehokkaat hakutyökalut, mallinnustyökalut • passiivinen: **suosittelujärjestelmät,** neuvonantajat, webpohjaiset näyteikkunat • yhdistelmä: asiakkaiden arvostelut, nettiyhteisöt	• muutoksia kuluttajatottumuksissa ja kysyntärakenteissa • positiivinen palaute nichetuotteiden osalta • kulttuurimuutokset yhä laajempaan tuotevalikoimaan pääsyn johdosta

4.2 Suosittelujärjestelmän valitseminen ja tietojen näyttäminen

Kuten aiemmin on mainittu, verkkokaupat ovat tuotteiden lukumäärän osalta niin suuria, että asiakkaat turhautuvat helposti verkkokaupassa asioidessaan. Suosittelujärjestelmät ovat apu tuotteiden parempaan ja tehokkaampaan näyttämiseen. Suosittelujärjestelmien etuna ovat erilaiset tavat näyttää tuotteita asiakkaille. Schafer ym. (1999) on huomannut yhtäläisyyksiä suosittelujen näyttämisessä perinteisen kaupankäynnin toimintatapojen kanssa. Moni näistä asioista toimii tehokkaammin. Esimerkiksi tuotetietämys ei rajoitu myyjän omaan tietämykseen. Tällaiset asiat voidaan nähdä merkittävänä etuna, mitä verkkokaupat tarvitsevat.

Suosittelujärjestelmää valitessa on otettava huomioon lukuisia eri asioita eivätkä kaikki suodatusmenetelmät sovellu kaikkiin tarpeisiin. Viimeisen vuosikymmenen aikana on kehitelty paljon uusia suodatusalgoritmeja (Ricci ym., 2011). Usein nämä tutkimukset vertaavat kehittelemäänsä algoritmiaan aiempiin esimerkiksi suorituskyvyssä ja suosittelujen tarkkuudessa. Suosittelujärjestelmää kehittävän tahon on perehdyttävä tarkoin, mitä algoritmeja hän haluaa käyttää palvelussaan. Ominaisuuksien ja toimintojen lisääminen jälkikäteen on kallista tai jopa mahdotonta.

Paaso (2012) luettelee tutkielmassaan kahdeksan lähtökohtaa suosittelujärjestelmän suunnitteluun:

- käytettävyys
- osallistuvuus
- sosiaalisuus
- välittömän mielihyvän tuottavuus
- käyttäjän itsenäisyyden jalostavuus
- nopea kehittyvyys
- navigoitavuus
- tasapainoisuus.

Verkkokaupan suosittelujärjestelmää määriteltäessä on huomioitava, millä tavoin ja mitä kanavia pitkin suositteluja halutaan asiakkaalle näyttää. Kilpailun asiakkaista ollessa suurta, on ensikokemuksen oltava mahdollisimman hyvä, jotta asiakas jäisi tutkimaan verkkokauppaa tarkemmin (Schafer ym., 2001).

Schafer ym. (1999) selittävät seitsemän menetelmää suosittelujen näyttämiseen asiakkaalle. Osa näistä ovat vanhoja menetelmiä perinteisestä liiketoiminnasta, jotka kuitenkin muokattuna versiona pätevät myös verkkoliiketoiminnassa ja suosittelujärjestelmien kontekstissa.

Selailu
Selailulla viitataan, perinteisessä kaupankäynnissä, tilannetta, missä asiakas kysyy kaupan myyjältä tuotetta liittyen tiettyyn kategoriaan. Asiakas alkaa selaamaan myyjän ehdottamia tuotteita ja valitsee itselleen mieluisimman. Tällöin on myyjän ammattitaidosta ja osaamisesta kiinni, mitä hän on asiakkaalle osannut suositella. Verkkoliiketoiminnassa suosittelujärjestelmää voidaan muokata manuaalisesti yhdistämällä tuotteisiin usean myyjän tai alan ammattilaisen suositukset ja näille voidaan määritellä tietyt painoarvot. Tämä edesauttaa tekemään selailijoista ostajia.

Vastaavat tuotteet
Yksi suosittelujärjestelmien ominaisuuksista on muistuttaa asiakasta vastaavista tuotteista. Tuote voi olla sellainen, minkä olemassaolon asiakas on unohtanut tai asiakkaalle täysin uusi tuote. Vastaavien tuotteiden lista muodostuu suosittelujärjestelmän seuratessa asiakkaan toimintaa joko implisiittisesti tai eksplisiittisesti. Implisiittisessä seurannassa seurataan asiakkaan selailua, eksplisiittisessä seurannassa asiakas on tehnyt joitain valintoja, esimerkiksi lisännyt tuotteen ostoskoriin. Tavoitteena tässä menetelmässä on ristikkäismyynti ja sitä kautta yksittäisen tilaussumman kasvattaminen.

Sähköposti
Hieman perinteistä suoramarkkinointia muistuttava menetelmä on lähettää asiakkaalle sähköpostia tuotteista, joista hän todennäköisesti on kiinnostunut. Yleensä sähköpostimuistutukset vaativat usein käyttäjän suostumuksen ja nii-

hin saa erilaisia manuaalisesti määriteltäviä ominaisuuksia, kuten ilmoituksen saamisen tuotteesta, joka on juuri saapunut myyntiin.

Tuotteiden kommentointi
Verkkokauppojen suosittelujärjestelmät osaavat antaa asiakkaalle suosituksia perustuen tuotteen saamiin muiden asiakkaiden tekemiin kommentteihin. Kommentteja käyttäen voidaan etsiä tiettyjä sanoja tai lauseita, joiden perusteella suosittelut muodostuvat. Ansaintamielessä tämä tekee selailijoista ostajia ja lisää lojaaliutta verkkokauppaa kohtaan.

Arvosana
Kuten luvussa 4 on usein mainittu, monet suosittelujärjestelmät perustuvat osittain tai kokonaan muiden käyttäjien antamiin arvosanoihin. Tuotteen arvottaminen voidaan toteuttaa joko jollain tietyllä asteikolla (esim. Likert) tai binäärisesti pidän/en pidä-menetelmällä. Hyvän keskiarvon muilta käyttäjiltä saanut tuote voi olla se tarvittava kannuste ostopäätökseen. Mahdollisuudessa antaa arvosana tuotteille on samat edut kuin kommentoinnissakin.

Top-N
Nimellä viitataan suosittelujärjestelmän tuottamaan personoituun listaan, joka on luotu analysoiden käyttäjän tottumuksia ja mieltymyksiä. Tuotteen, jota asiakas on harkinnut ostavan, näkeminen suosittelujärjestelmän luomalla listalla voi kannustaa asiakasta ostopäätökseen.

Järjestellyt hakutulokset
Hieman Top-N-menetelmää muistuttaen hakutuloksia voidaan järjestää perustuen käyttäjästä kerättyihin tietoihin. Top-N menetelmässä näytettävät tuotteet ovat rajoitettu johonkin lukumäärään (esim. ”Kymmenen sopivinta tuotetta Teille”) hakutulosten jatkuessa laajemmalle.

5 Yhteenveto

Tässä tutkielmassa käsiteltiin big data-analyysia suosittelujärjestelmien kautta. Luku kaksi selitti big datan ja suosittelujärjestelmien välisen yhteyden. Myöhemmät luvut keskittyivät yksinomaan suosittelujärjestelmiin, niiden tekniseen toteutukseen sekä niiden tarjoamiin hyötyihin.

Tutkimusaihe osoittautui melko laajaksi, joka tarkoittaa usean luvun jättämistä hieman vajavaiseksi. Tutkielmia, jotka sisältävät termin yhteisöllinen suodatus, löytyy Google Scholar-hakupalvelusta 320 000 kappaletta. Toisin sanoen koko tutkielman aihe olisi voinut käsitellä vain yhtä suodatustapaa ja olisi silti jättänyt useat tärkeät asiat esittelemättä.

Toisaalta tämän tutkielman tarkoitus on olla yleiskatsaus markkinoilla oleviin verkkoliiketoiminnassa käytettäviin suosittelujärjestelmiin, eikä niinkään kaiken kattava eepos. Taustakirjallisuudesta on poimittu keskeisimmät seikat kuvaamaan sekä suosittelujärjestelmiä että niiden antamia hyötyjä. Tutkielman laajuutta joutui rajoittamaan sekä työkalujen että hyötyjen kuvauksessa, mutta oleellisimmat asiat on saatu mukaan.

5.1 Tulokset ja johtopäätökset

Tutkielmassa tehdyt havainnot antoivat tutkimuskysymyksiin melko kattavan vastauksen. Työkalut suosittelujärjestelmän taustalla- kysymykseen vastattiin esittelemällä neljä erilaista suosittelujärjestelmän variaatiota sekä hybridin vaihtoehdon. Yhteisöllinen, sisältöpohjainen, demografinen, tietämyspohjainen ja hybridit menetelmät kattavat suuren osan verkkoliiketoiminnassa käytettävistä suosittelujärjestelmistä. Tässä ei kuitenkaan ole kaikki mahdolliset suodatustavat, vaan uusia algoritmeja ja suodatustapoja syntyy jatkuvasti lisää. Suosittelujärjestelmät ovat suuressa suosiossa kaupallisten palveluiden lisäksi myös tutkimusten kohteena, jonka ansiosta uusia tekniikoita syntyy säännöllisesti.

Uudet algoritmit ja suodatustekniikat suosittelujärjestelmissä useimmiten perustuvat kuitenkin jo aiemmin hyväksi todettuihin tekniikoihin, usein joko

CF:iin tai CBF:iin. Myös erilaisia hybridiratkaisuja on lukuisia. Hybrideistä kertova luku on tässä tutkielmassa melko rajallinen selittäen seitsemän eri variaatiota. Toisaalta useamman variaation käsitteleminen ei olisi järkevää eikä mahdollista, tutkielman laajuus huomioon ottaen. Suosittelujärjestelmiä esitellessä käsiteltiin jokaisen kohdalla vahvuudet ja heikkoudet. Tämä auttaa havainnollistamaan niiden käyttötarkoitusta. Tässä tutkielmassa esitetyt suodatustekniikat, hybridi tekniikka poislukien, on esitetty vahvuuksineen ja heikkouksineen taulukossa 5. Taulukko 5:n sisältämät vahvuudet ja heikkoudet ovat yhteenveto luvun 3 tekstisisällöstä. Hybridi suodatustekniikka on käsitteenä niin laaja, että sen vahvuudet ja heikkoudet eivät ole yleistettävissä. Hybridin suodatustekniikan ideana on räätälöity kokonaisuus, joka minimoi suunnitellussa kohteessa ilmentyvät kriittiset heikkoudet.

TAULUKKO 5 Suodatustekniikoiden vahvuudet ja heikkoudet

Suodatustekniikka	Vahvuudet	Heikkoudet
Yhteisöllinen suodatus	• helppo käyttöönotto • ei tarvitse kattavaa tuotetietokantaa • mukautuvuus eri sisältötyyppeihin • kykenee tarjoamaan yllätyksellistä sisältöä	• kylmäkäynnistysongelma • ei mahdollista selittää suosituksia, koska ne perustuvat vertaisten toimintaan. Ts. huono läpinäkyvyys • huono skaalautuvuus (muistipohjainen CF)
Yhteistoimintapohjainen suodatus	• riippumattomuus muista käyttäjistä • läpinäkyvyys • ei kylmäkäynnistysongelmaa uuden tuotteen kohdalla	• vaatii tarkan tuotetietokannan • yli-erikoistuminen • kylmäkäynnistysongelma koskien uutta käyttäjää • ei niin tarkka kuin CF
Demografinen suodatus	• ei kylmäkäynnistysongelmaa, mikäli laaja demografinen käyttäjäprofiili saatavilla • nopea käyttöönotto	• rajallinen käyttöalue. Heikko mukautuvuus eri sisältötyyppeihin.
Sisältöpohjainen suodatus	• osaa suositella samaan alueeseen kuuluvia muita tuotteita (esim. hakutoimintoa käytettäessä) • ei vaadi käyttäjän tai vertaisten tuotearvioita toimiakseen • ei kylmäkäynnistysongelmaa	• vaatii toimialuetietämystä ja kattavan tuotetietokannan • koneoppimiskomponentit tehokkaan toiminnan edellytyksenä

Luvun 4 tarkoitus on vastata jälkimmäiseen tutkimuskysymykseen: Mitä hyötyjä verkkokauppojen suosittelujärjestelmät tarjoavat palveluntarjoajalle?

Suosittelujärjestelmien on havaittu parantavan konversiota (kuinka hyvin jokin tietty sivusto saavuttaa tavoitteensa suhteessa käyntimäärään), nostavan keskimääräistä myyntimäärää ja nopeuttavan tuotteiden löytämistä. Tärkeimpänä hyötynä nähtäneen parantunut asiakastyytyväisyys ja sitä kautta suurempi konversioprosentti. Aiempi tutkielmakirjallisuus kartoitti taloudellisia hyötyjä yleisestä näkökulmasta. Useat tutkimukset mainitsivat suosittelujärjestelmien antavan vastaavanlaisia hyötyjä, joten tietoa voi siltä osin pitää luotettavana. Jos aiemmasta kirjallisuudesta löytyi spesifimpää tietoa suosittelujärjestelmien antamista hyödyistä, koskivat ne jonkin yksittäisen palveluntarjoajan suosittelujärjestelmää. Tämän kaltaisten tietojen mukaan ottaminen heikentäisi tutkimustulosten yleistettävyyttä.

Luku 4.1. koskien pitkä häntä-ilmiötä verkkokaupassa on huomionarvoinen asia puhuttaessa suosittelujärjestelmän eduista. Pitkä häntä toteutuu, mikäli kaupassa on todella laaja tuotetarjonta. Käytännössä katsoen kyseessä on oltava verkkokauppa ja liiketoimintamallia on suotavaa muokata pitkä häntäilmiötä suosivaksi. Laajan tuotevalikoiman lisäksi tuotteet on saatava asiakkaan nähtäväksi eikä tämä onnistu ilman suosittelujärjestelmiä. Tätä asiaa on käsitelty yllättävän vähän aiemmassa kirjallisuudessa. Pitkästä hännästä ja suosittelujärjestelmistä on muutama erillinen tutkielma, esimerkiksi Brynjolfsson ym. (2006) tutkielma From niches to riches: Anatomy of the long tail. Tämän luvun keskeisin lähde oli Chris Andersonin kirja Pitkä häntä: Miksi tulevaisuudessa myydään vähemmän enempää (2006).

Luku 4.2. kertoo pintapuolisesti, mitä on otettava huomioon suosittelujärjestelmää valitessa. Tämä on yksi esimerkki aihepiiristä, johon olisi voinut keskittyä erittäin syvällisesti, kenties erillisen empiirisen tutkielman verran. Tästä aihepiiristä on tehty aiempia tutkimuksia melko rajallisesti.

5.2 Pohdittavaa ja mahdollisia jatkotutkimusaiheita

Tässä tutkielmassa ei ole otettu huomioon sosiaalisen median merkitystä suosittelujärjestelmissä. Sen merkitys näkyy muun muassa siinä, että yhä useampi asiakas päätyy verkkokauppaan sosiaalisessa mediassa nähdyn mainoksen perusteella. Sosiaaliseen mediaan perustuva suosittelujärjestelmät pohjautuu usein ihmisiin ja ns. tägeihin (engl tag). Ihmisten, tägien ja tuotteiden välistä informaatiota voidaan kerätä erilaisista sosiaalisen median kanavista, kuten blogeista, kirjanmerkeistä, yhteisöistä, wiki-sivuista ja jaetuista tiedostoista (Guy, Zwerdling, Ronen, Carmel & Uziel, 2010)

Yksi alati kasvava alue on mobiilialustat suosittelujärjestelmissä. Verkkokauppoja käytetään kasvavissa määrin mobiililaitteilla, kuten älypuhelimilla. Näissä alustoissa on huomioitava näytön rajallinen koko ja kosketusnäytön vaatimat ominaisuudet käyttöliittymää suunnitellessa. Varsinaisten ostotapahtumien lisäksi mobiililaitteita käytetään oston tukena – tuotteita ja verkkokauppoja selataan älypuhelimella, mutta varsinainen ostos voidaan tehdä tietokoneella.

Viimeisimpänä tuotekategoriana mobiiliteknologiassa on älykellot. Miten älykellojen tuottama data saadaan jalostettua suosittelujärjestelmiin?

Suosittelujärjestelmiä sovelletaan useissa erilaisissa palveluissa ja sivustoissa, ei pelkästään verkkokaupoissa. Aiemmin mainitsemani sosiaaliseen mediaan perustuva suodatus on luonnollisesti yleinen menetelmä erilaisissa sosiaalisen median palveluissa. Suosittelujärjestelmiä voidaan käyttää myös työkaluna yrityksen rekrytointiprosessissa, mistä Paaso (2012) onkin tehnyt tutkielman. Sovellettavia kohdealueita on lukuisia, mikä takaa suosittelujärjestelmien suosion tutkimusaiheena pitkäksi ajaksi eteenpäin.

LÄHTEET

Anand, S.S., Mobasher, B. (2005). Intelligent techniques for web personalization. In: Intelligent Techniques for Web Personalization, *Springer 1–36.*

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more.* Hyperion.

Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of librar y and information systems 69 (Supplement 32*), 175-186.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction 12 (4), 331-370.*

Burke, R. (2007). Hybrid web recommender systems. *Teoksessa The adaptive web. Springer, 377-408.*

Brynjolfsson, E., Hu, Y. & Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the valueof increased product variety at online booksellers. *Management Science 49 (11*), 1580-1596.

Brynjolfsson, E., Hu, Y. J. & Smith, M. D. (2006). From niches to riches: Anatomy of the long tail. *Sloan management review 47 (4), 67-71.*

Franks, B. (2012). Wiley and SAS Business Series : Taming the Big Data Tidal W ave : Finding Opportunities in Huge Data Streams with Advanced Analytics. *Hoboken, NJ, USA: Wiley*.

Givon, S. (2011). Predicting and using social tags to improve the accuracy and tr ansparency of recommender systems. *The University of Edinburgh*

Guy, I., Zwerdling, N., Ronen, I., Carmel, D., Uziel, E. (2010). Social media recommendation based on people and tags. *ACM Transactions on Information Systems.*

Herlocker, J., Konstan, J., Riedl, J. (2000). Explaining collaborative filtering recommendations. *ACM.*

Herlocker, J. L., Konstan, J. A., Terveen, L. G. & Riedl, J. T. (2004). Evaluating co llaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS) 22 (1),* 5-53.

Konstan, J. A. & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction 22 (1-2), 101-123.*

Jian, J. & Qun C. (2012). A trust-based Top-K recommender system using social tagging network. *Fuzzy Systems and Knowledge Discovery, 2012 9th International Conference.*

Paaso, M. (2012). Suosittelujärjestelmät rekrytointiprosessissa. *Oulun Yliopisto*

Ricci, F., Rokach, L., Shapira, B., Kantor, P. (2011). *Recommender systems handbook*. Springer US.

Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. (2001). Item-based collaborative fi ltering recommendation algorithms. *Proceedings of the 10th international conference on World Wide Web. ACM*, 285.

Schafer, J. B., Frankowski, D., Herlocker, J. & Sen, S.(2007). Collaborative filterin g recommender systems. *Springer, 291-324.*

Schafer, J. B., Konstan, J. A. & Riedl, J. (2001). E-commerce recommendation applications. *In Applications of Data Mining to Electronic Commerce. Springer, 115-153.*

KAUPALLISET LÄHTEET

Laney, D., (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety. Haettu 19.2.2015 osoitteesta http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

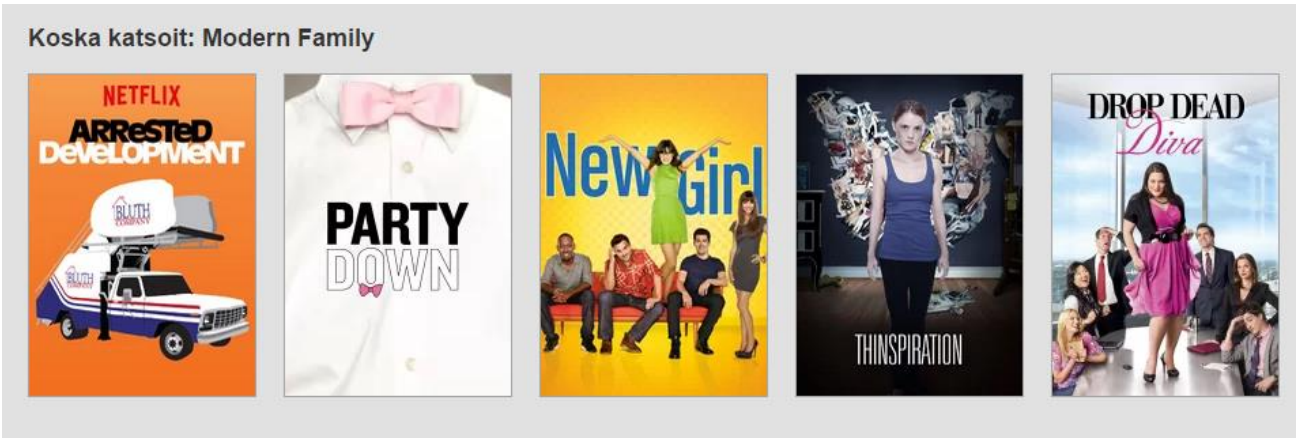
Intel. (2013). A vision from big data. Haettu 19.2.2015 osoitteesta: http://www.intel.com/content/dam/www/public/us/en/documents/reports/intel-corp-big-data-policy-position-paper.pdf

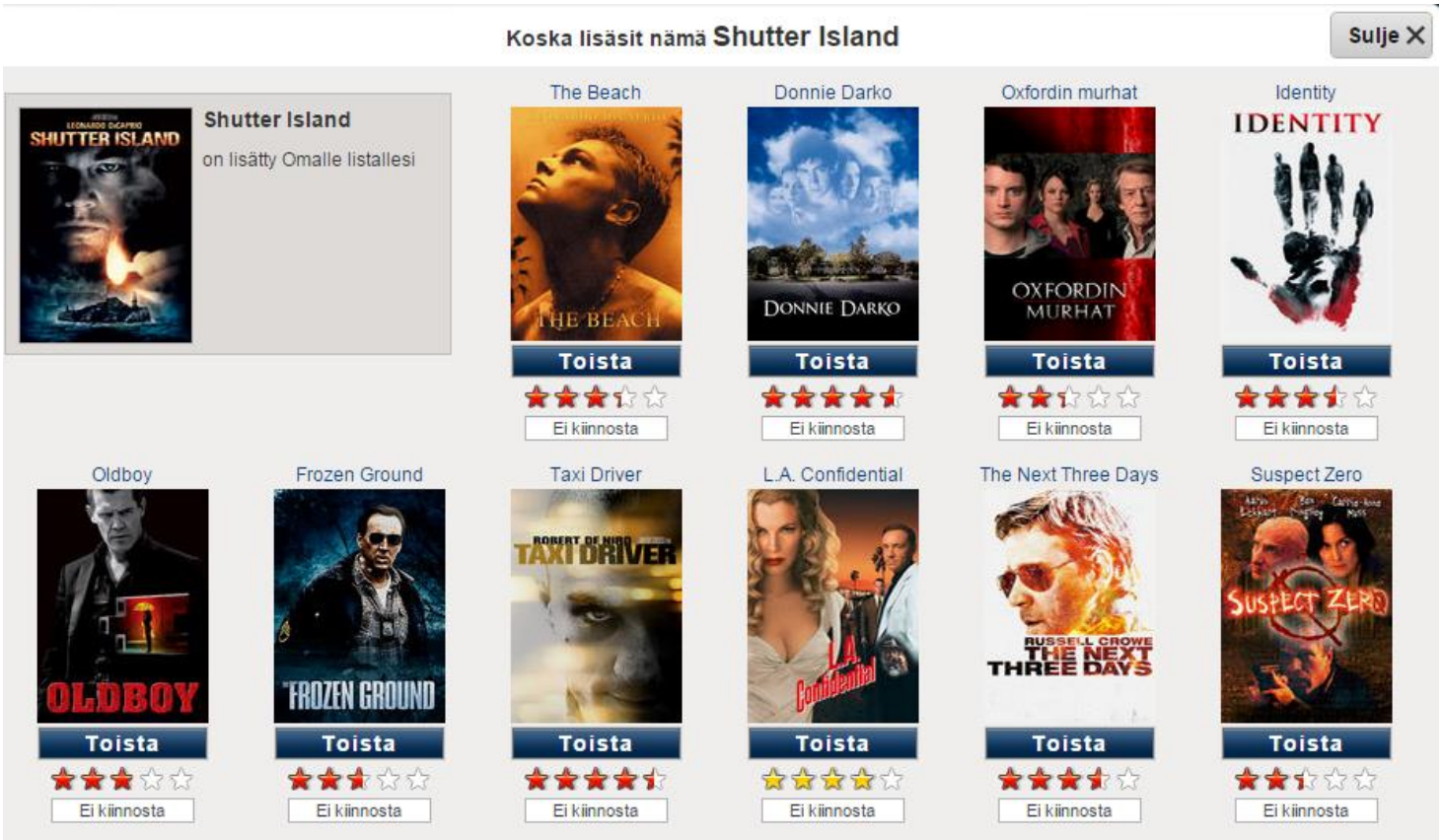
Techcrunch (2010). Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003 Haettu 19.2.2015 osoitteesta: http://techcrunch.com/2010/08/04/schmidt-data/

Netflix. Liitetiedostot haetty 3.3.2015 osoitteesta: www.netflix.com

LIITE 1 YHTEISÖLLINEN SUODATUS NETFLIXISSÄ

LIITE 2 YHTEISTOIMINTAPOHJAINEN SUODATUS NETFLIXISSÄ





LIITE 3 TIETÄMYSPOHJAINEN SUODATUS NETFLIX-PALVELUN HAKU-TOIMINNOSSA