TAMPEREEN TEKNILLINEN YLIOPISTO

JUHA HAAVISTO

SOSIAALISET PIIRTEET VERKKOPALVELUN SUOSITTELUJÄRJESTELMISSÄ

Diplomityö

Tarkastajat: tutkija Jukka Huhtamäki, dosentti Ossi Nykänen
Tarkastajat ja aihe hyväksytty Tieto- ja sähkötekniikan tiedekunta-neuvoston kokouksessa 3. maalis-kuuta 2010

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Tietotekniikan koulutusohjelma
HAAVISTO, JUHA: Sosiaaliset piirteet verkkopalvelun suosittelujärjestelmissä
Diplomityö, **60** sivua, **2** liitesivua
Maaliskuu 2012
Pääaine: Hypermedia
Tarkastajat: tutkija Jukka Huhtamäki, dosentti Ossi Nykänen
Avainsanat: Suosittelujärjestelmä, sosiaalinen media, sosiaaliset verkostot, sosiaaliset verkostoitumispalvelut, yhteisöllinen suodatus, CF, sosiaalinen suosittelu

Internetissä kuluttajan saatavilla on nykyään valtava, koko ajan kasvava määrä tuotteita ja palveluja. Erilaiset tiedonhaku-, suodatus- ja suosittelumenetelmät auttavat kuluttajaa löytämään haluamansa tuotteen. Erityisesti suosittelujärjestelmien rooli korostuu tarjotessa käyttäjälle uusia elämyksiä tai ennen näkemättömiä tuotteita. Suosituimpia tapoja toteuttaa suosittelujärjestelmä on yhteisöllinen suodatus. Se perustuu suositeltavan tuotteen sisällön analysoinnin sijaan järjestelmän käyttäjien arvioihin eri tuotteista.

Web-teknologioiden kehitys on tuonut yhteisöllisyyden ja käyttäjälähtöisen näkökulman osaksi nykypäivän verkkopalveluja. Sosiaalisen median palveluista suosiota ovat keränneet erityisesti sosiaaliset verkostoitumispalvelut, joiden voima perustuu yhteisön luomaan sisältöön ja lisäarvoon. Sosiaaliset verkostot kehittyvät yleensä jonkin kiinnostusaiheen ympärille tai keskittyvät tosielämän sosiaalisten suhteiden mallintamiseen.

Tässä tutkielmassa esitetään tapoja hyödyntää käyttäjän sosiaalista verkostoa automaattisissa suosittelujärjestelmissä. Tutkimusten mukaan käyttäjät luottavat suositusten saamiseen enemmän ystäviinsä kuin automaattisiin suosittelujärjestelmiin. Hyödyntämällä olemassa olevaa sosiaalista verkostoa voidaan parantaa suosittelutuloksia ja lisätä luottamusta saatuihin tuloksiin. Työn tarkoituksena on tarkastella sosiaalisen verkoston lisäarvoa suositteluissa ja menetelmän yleiskäyttöisyyttä.

Sosiaalisen verkoston vaikutusta selvitettiin simuloimalla järjestelmää, joka hyödyntää sosiaalista verkostoa suositteluun. Tutkimuksessa hyödynnettiin aineistoa olemassa olevasta, yhteisöllistä suodatusta käyttävästä suosittelujärjestelmästä. Sosiaaliseen verkostoitumispalveluun kohdistuneella kyselyllä verrattiin käyttäjien arvioita olemassa olevaan dataan. Havainnot osoittivat, että suosittelutulokset paranivat verrattuna oletustoteutukseen. Vaikka saatuja tuloksia ei voida suoraan vertailla, vahvistavat ne kuitenkin aikaisempien tutkimuksien havaintoja. Idean jatkokehittelyllä onkin mahdollista saada aikaan parempia tuloksia.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Information Technology
HAAVISTO, JUHA: Social Features in a Recommender System
Master of Science Thesis, **60** pages, **2** Appendix pages
March 2012
Major: Hypermedia
Examiners: Researcher Jukka Huhtamäki, Adjunct Professor Ossi Nykänen
Keywords: recommender system, social media, social networks, social networking sites, collaborative filtering, CF, social filtering

A user browsing the internet today faces a vast amount of products and services offered, which is also growing daily. Different information retrieval and recommendation methods help the user to find relevant or interesting content. Although many implementations for recommender systems exist, the most popular method is called collaborative filtering. Instead of analyzing the content of the recommended product, recommendations are based on the opinions of users with similar taste.

Advances in web technologies have brought a communal and a user-centric view to modern websites. Social media sites – especially social networking sites – have proven popular. The strength of these services is based on user-created content and social networking. Networks form usually around users' subjects of interest. They might also model the real-life social networks of people.

The goal of this study is to present different methods of using a user's social network in an automatic recommender system. Studies show that users prefer recommendations that are based on the opinions of people they know. By utilizing a user's existing social network, it is possible to provide more accurate and trustworthy recommendations. The purpose is to evaluate social recommendation methods and provide practical suggestions for implementation.

The results for social recommendation were simulated by fusing existing data from a collaborative filtering system with data from a social networking site. Ratings from the social networking website were gathered using a questionnaire and then merged with existing data. The results showed noticeable improvement over the default implementation. Even though these results are not comparable, similar results have been gathered from previous studies. These show that fusing social data with collaborative filtering does indeed yield better recommendation results.

ALKUSANAT

Tutkielma sai alkunsa maaliskuussa 2010, jolloin hyväksytin aiheen tiedekunnan kokouksessa. Aiheen valintaan vaikutti kiinnostus suosittelujärjestelmän käytännön toteutukseen. Työn oli alun perin tarkoitus keskittyä enemmänkin perinteisten suosittelujärjestelmien vertaamisen ja suosittelujärjestelmän toteutukseen osana erästä verkkopalvelua. Aihe jäi lopulta hautumaan pitkäksi aikaa, kunnes syksyllä 2011 tartuin työhön jälleen uuden näkökulman myötä.

Haluan kiittää työni tarkastajia tutkija Jukka Huhtamäkeä ja dosentti Ossi Nykästä arvokkaasta palautteesta ja asiantuntevasta ohjauksesta työn suhteen. Työ on omistettu isäni Rauno Haaviston muistolle.

Tampereella 29. helmikuuta 2012

Juha Haavisto

SISÄLLYS

1 Johdanto 1
1.1 Taustaa 1
1.2 Työn tavoite ja tutkimuskysymykset 2
1.3 Työn rakenne 3
2 Suosittelujärjestelmät 4
2.1 Sisältöpohjainen suodatus 6
2.2 Yhteisöllinen suodatus 7
2.2.1 Muistipohjaiset suodatusmenetelmät 12
2.2.2 Mallipohjaiset suodatusmenetelmät 15
2.3 Muita suosittelumenetelmiä 15
2.4 Suosittelujärjestelmien ongelmat ja haasteet 16
2.5 Suosittelujärjestelmän arviointi 17
2.5.1 Määrällinen arviointi 17
2.5.2 Laadullinen arviointi 19
3 Sosiaalinen media ja sosiaaliset verkostot 21
3.1 Sosiaaliset verkostoitumispalvelut 23
3.1.1 Facebook 24
3.1.2 Google+ 24
3.1.3 Last.fm 24
3.1.4 LinkedIn 25
3.2 Sosiaalisten verkostojen teoriaa 25
3.3 Liittymät sosiaalisiin verkostoitumispalveluihin 28
3.4 Sosiaalinen suosittelu 30
4 Sosiaalisen verkoston hyödyntäminen yhteisöllisessä suodatuksessa 32
4.1 Tutkimusasetelma 32
4.2 Toteutustekniikat 34
4.2.1 Relaatiotietokanta ja MySQL 34
4.2.2 PHP 37
4.2.3 JSON 37
4.3 Aineisto 37
4.3.1 MovieLens-palvelun aineisto 38
4.3.2 Kyselytutkimuksen tulokset 40
4.3.3 Facebook-käyttäjien tiedot 41
4.3.4 Koottu aineisto 43
4.4 Sosiaalisen verkoston painotus 45
4.5 Tutkimuksen tulokset 46
4.5.1 Käyttäjien samankaltaisuus 46
4.5.2 Suosittelutulokset 48
5 Johtopäätökset ja pohdinta 51
5.1 Toteutuksen helppous 52

5.2 Suosittelutulosten paraneminen 53
5.3 Suosittelujärjestelmien ongelmien kompensointi 54
5.4 Jatkokehitysideat 54
6 Yhteenveto 56
Lähteet 57
Liitteet 61

KUVAT

Kuva 2.1. Yhteisöllisen suosittelujärjestelmän toiminta. 4
Kuva 2.2. Suosittelumenetelmien jaottelu. 6
Kuva 2.3. Yhteisöllisen suodatuksen prosessi. 9
Kuva 2.4. Samankaltaisten käyttäjien valinta naapurustolla N = 3. 12
Kuva 3.1. Sosiaalisen median elementit (Kangas et al. 2007). 22
Kuva 3.2. Suuntaamattoman verkoston sosiogrammi. 27
Kuva 3.3. Facebookin Graph API -rajapinnan käyttö (Tamada 2011). 29
Kuva 4.1. Tutkimuksen toteutusasetelma. 33
Kuva 4.2. Tutkimusaineisto. 38
Kuva 4.3. Ote kyselytutkimuksen lomakkeesta. 41
Kuva 4.4. Eri palveluista koottu tutkimusaineisto. 43
Kuva 4.5. Arvosanajakaumat MovieLens- ja IMDB-palveluista sekä kyselytutkimuksesta. 44
Kuva 4.6. Facebook-kontaktien määrät samankaltaisten käyttäjien joukossa eri naapurustoilla. 47
Kuva 4.7. Sosiaalisen verkoston vaikutus suosittelutuloksiin, kun $\lambda = 0.9$ ja $\lambda = 0.8$. 49
Kuva 4.8. Sosiaalisen verkoston vaikutus suosittelutuloksiin, kun $\lambda = 0.1 \ldots 0.7$ 50

TAULUKOT

Taulukko 2.1. Esimerkki arvostelumatriisista **R**. 10
Taulukko 2.2. Esimerkki käyttäjien samankaltaisuusmatriisista $\mathbf{S}_u$. 11
Taulukko 2.3. Esimerkki tuotteiden samankaltaisuusmatriisista $\mathbf{S}_i$. 11
Taulukko 2.4. Esimerkki ennustematriisista **P**. 14
Taulukko 3.1. Kaksiarvoinen (a) ja arvotettu (b) sosiomatriisi. 26
Taulukko 4.1. Käyttäjärelaatio. 35
Taulukko 4.2. Käyttäjien samankaltaisuus -relaatio. 36
Taulukko 4.3. Elokuvarelaatio. 36
Taulukko 4.4. Arvostelurelaatio. 36
Taulukko 4.5. MovieLens 1M -tietosarja. 39
Taulukko 4.6. Tutkimusaineiston koko. 44

KÄYTETYT LYHENTEET

Apache	Vapaaseen lähdekoodiin perustuva palvelinohjelmisto
API	Application Programming Interface, ohjelmointirajapinta
FB	Facebook, suosittu sosiaalinen verkostoitumispalvelu
HTML	Hypertext Markup Language, www-sivuilla käytetty merkkauskieli
HTTP	Hypertext Transfer Protocol, internetissä käytetty tiedonsiirtoprotokolla
IMDB	Internet Movie Database, suosittu elokuvasivusto
JavaScript	Selaimessa tulkattava skriptikieli
JSON	JavaScript Object Notation, rakenteinen datan esitysmuoto
MAE	Mean Absolute Error, absoluuttinen keskimääräinen virhe
MySQL	Vapaaseen lähdekoodiin perustuva relaatiotietokanta
PHP	PHP: Hypertext Preprocessor, tulkattava ohjelmointikieli
RMSE	Root Mean Square Error, keskineliövirheen neliöjuuri
SNA	Social Network Analysis, sosiaalisten verkostojen analyysi
SNS	Social Networking Sites, sosiaaliset verkostoitumispalvelut
SQL	Structured Query Language, rakenteinen kyselykieli
WBSN	Web-Based Social Network, web-pohjainen sosiaalinen verkosto
web	Ks. www
www	World Wide Web, internetissä käytetty palvelujärjestelmä

1 JOHDANTO

Suosittelujärjestelmiä on käytössä useissa internetin verkkokaupoissa, joten kaupallinen tarve ruokkii tarvetta parempien suosittelujen saamiseksi. Suosittelujärjestelmiä on tutkittu jo lähes kaksivuosikymmentä, ja uusia algoritmillisia ratkaisuja suositusten parantamiseksi on tarjolla runsaasti. Viime vuosina painopiste on kuitenkin siirtynyt pelkästä algoritmien optimoinnista muidenkin keinojen etsintään. Suosituksia voidaan yrittää parantaa esimerkiksi sosiaalisia tai luottamusverkostoja hyödyntämällä. Tässä tutkielmassa pohditaankin, voidaanko suosittelutuloksia parantaa sosiaalisten suhteiden avulla.

Seuraavissa aliluvussa esitetään työn tausta, tutkimuskysymykset ja työn rakenne.

1.1 Taustaa

Internet on pienentänyt maailmaamme tuomalla palveluita ja tuotteita entistä lähemmäksi kuluttajaa. Tuskin koskaan on kuluttajalla ollut näin paljon tietoa saatavilla samaan aikaan. Toisaalta tiedon paljous aiheuttaa myös ongelmia, sillä valinnanvaran määrä on usein häkellyttävä. Lisäksi internetissä tarjolla olevan tiedon määrä kasvaa koko ajan. Miten kuluttaja voi löytää esimerkiksi haluamansa tuotteen verkkokaupan kirjaimellisesti tuhansista toisiaan muistuttavista tuotteista?

Ongelma on tuttu monelle verkkokaupalle. On mahdotonta saada kaupan koko valikoimaa käyttäjän näkyville, eikä siinä olisi edes mitään mieltä. Esimerkiksi verkkokauppa Amazonin[1] listoilla oli jo vuosituhannen vaihteessa yli 18 miljoonaa tuotetta (Dé 2000). Verkkopalvelun on mahdotonta tuntea käyttäjäänsä perinpohjin, jolloin se voisi suoraan suositella hänelle esimerkiksi mielenkiintoista kirjaa. Tähän tarpeeseen vastaavat erilaiset suosittelujärjestelmät, jotka pyrkivät tarjoamaan käyttäjää kiinnostavia kohteita tarjolla olevasta valikoimasta.

Suosittelujärjestelmät soveltavat käyttäjien liikkeistä louhittua tietoa ja pyrkivät tekemään personoituja suosituksia jokaiselle käyttäjälle (Sanastokeskus 2010). Suositeltu tieto voi liittyä tuotteisiin, informaatioon tai palveluihin[2]. Erityisesti samankaltaisten naapureiden etsintään perustuvat suosittelujärjestelmät ovat saavuttaneet suuren suosion www-sivustoilla (Symeonidis et al. 2006). Ensimmäinen suosittelujärjestelmä esiteltiin vuonna 1992 ja ensimmäinen tunnettu sovellutus oli GroupLens-järjestelmä vuodelta 1994 (Goldberg et al. 1992, Resnick et al. 1994).

[1] http://www.amazon.com
[2] Jatkossa suositeltavista asioista käytetään yksinkertaisuuden vuoksi vain nimitystä *tuote*.

Lopullinen läpimurto tapahtui Amazonin myötä, jonka yhteisölliseen suodatukseen perustuva suosittelujärjestelmä saavutti suurta suosiota. Yksinkertainen esimerkki tällaisesta suosittelusta on Amazonin ”tuotteet, joita muutkin ovat ostaneet” -listat.

Sanastokeskuksen (2010) määritelmän mukaan sosiaalinen media (*social media*) on ”tietoverkkoja ja tietotekniikkaa hyödyntävä viestinnän muoto, jossa käsitellään vuorovaikutteisesti ja käyttäjälähtöisesti tuotettua sisältöä ja luodaan ja ylläpidetään ihmisten välisiä suhteita”. Käyttäjien lisäämä sisältö voi olla käytännössä mitä vain, esimerkiksi keskustelua tai kuvien lisäystä ja niiden kommentointia. Yhteisöllisyyteen liittyen käyttäjillä on usein verkkopalveluun liittyvä oma profiili, jonka pohjalle käyttäjän sosiaalinen verkosto rakentuu (boyd & Heer 2006). Eräs nykyään suosittu verkkopalvelun muoto onkin juuri sosiaaliset verkostoitumispalvelut, jossa pääpaino on käyttäjien välisellä interaktiolla.

Perinteisesti suosittelujärjestelmät ovat hyödyntäneet niiden sisäisesti keräämää tietoa suosittelupäätöksen tekoon. Suosittelupäätöksessä voidaan myös hyödyntää käyttäjän sosiaalista verkostoa luottamuksen lisäämiseksi. Tunnetulta ihmiseltä saatu suosittelu saattaa olla käyttäjälle mieluisampi kuin koneen tarjoama, mutta tuttuus ei välttämättä takaa tuotteen sopivuutta omaan makuun (Lee & Brusilovsky 2009).

Vaikka erilaisten palveluiden yhdistäminen ei olekaan suoraan mahdollista, tarjoavat useat sosiaalisen median palvelut ja sosiaaliset verkostoitumispalvelut kehittäjien käyttöön rajapintoja, joilla niiden verkostoja voidaan ohjelmallisesti käsitellä. Sosiaalisen verkostotiedon ja yhteisöllistä suodatusta käyttävän suosittelujärjestelmän käyttäjätietojen yhdistäminen luo mielenkiintoisen hypoteesin – voidaanko suositusten laatua parantaa sillä, että suositus perustuu henkilöön, jonka käyttäjä tuntee oman sosiaalisen verkostonsa kautta?

1.2 Työn tavoite ja tutkimuskysymykset

Tutkielmassa tarkoituksena on selvittää, miten tietoa sosiaalisista verkostoista ja niiden suhteista voidaan hyödyntää suosittelujärjestelmässä. Kysymykseen pyritään saamaan vastaus tarkastelemalla sosiaalisia verkostoja sekä kokeilemalla erilaisia suositteluteknikoita. Erityisesti tutkielmassa keskitytään yhteisölliseen suodatukseen, sillä se on ylivertaisesti suosituin käytetyistä suosittelumenetelmistä. Työssä on kuitenkin myös esitelty muita tapoja.

Päätutkimuskysymys, johon tutkielmassa haetaan vastausta on:

> *Onko mahdollista hyödyntää käyttäjän sosiaalista verkostoa suosittelujärjestelmässä?*

Tämän tutkimuskysymyksen taustalla on hypoteesi, jonka mukaan tutuilta ihmisiltä saadut suositukset ovat parempia kuin tuntemattomilta käyttäjiltä tulevat. Tarkentavia ja pääkysymykseen vastaamista tukevia tarkentavia tutkimuskysymyksiä ovat:

1. *Kuinka helppoa on toteuttaa sosiaalista verkostoa hyödyntävä suosittelujärjestelmä?*
2. *Paranevatko suosittelutulokset tai suositusten laatu sosiaalista verkostoa hyödyntämällä?*
3. *Voidaanko sosiaalisten verkostojen avulla kompensoida suosittelujärjestelmien ongelmia?*

Ensimmäisessä tarkentavassa tutkimuskysymyksessä selvitetään suosittelujärjestelmän vaatimukset ja edellytykset, joiden pohjalle voidaan rakentaa sosiaalista verkostoa hyödyntävä osa. Kysymyksen kautta sosiaalisten piirteiden hyödyntämiselle voidaan muodostaa yksi osa siitä kontekstista, jossa sen on tarkoitus käytännössä toimia.

Toista tarkentavaa tutkimuskysymystä varten suoritetaan kyselytutkimus, minkä avulla mallinnetaan suosittelujärjestelmää, joka hyödyntää käyttäjän sosiaalista verkostoa. Etenkin testataan hypoteesia, jonka mukaan tutuilta saadut suositukset ovat parempia kuin nimettömän käyttäjäyhteisön arvioiden perusteella lasketut suositukset. Saatuja tuloksia arvioidaan sekä määrällisesti että laadullisesti. Tarkoitus ei ole kehittää mahdollisimman tehokasta suosittelualgoritmia vaan pohtia keinoja, joilla suositusten laatua voidaan parantaa.

Kolmannen tarkentavan tutkimuskysymyksen myötä pohditaan keinoja, joilla voidaan estää tai kompensoida suosittelujärjestelmissä yleisesti havaittuja ongelmia. Suosittelujärjestelmien ongelmista ja haasteista kerrotaan tarkemmin luvussa 2.4.

1.3 Työn rakenne

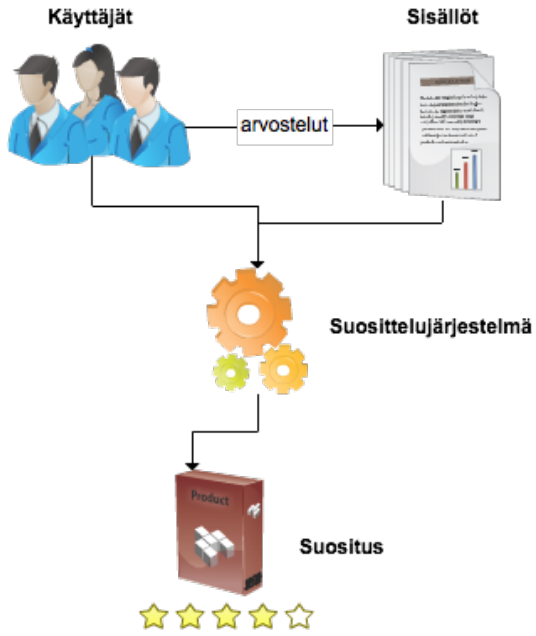
Tutkielma on jäsennelty seuraavasti. Toisessa luvussa esitetään suosittelujärjestelmien tarkoitus ja niihin liittyvät periaatteet. Tarkemmalla tasolla käydään läpi yhteisöllisen suodatuksen toimintaperiaate. Lisäksi esitellään suosittelujärjestelmien toteutuksessa yleensä eteen tulevia ongelmia sekä suosittelujärjestelmien arviointia. Kolmannessa luvussa luodaan katsaus sosiaaliseen mediaan, sosiaalisiin verkostoitumispalveluihin sekä sosiaalisiin verkostoihin. Sosiaalisista verkostoista esitellään niiden teoriaa yleistasolla ja esitetään työn kannalta olennaiset sosiaaliset verkostoitumispalvelut.

Neljännessä luvussa esitetään tutkimus ja sen toteutus, minkä avulla pyritään antamaan vastaus päätutkimuskysymykseen ja sen tarkentaviin kysymyksiin. Tutkimuksen idea, lähtöaineisto ja toteutuksen yksityiskohdat käydään läpi tässä luvussa. Lisäksi esitetään tutkimuksen tulokset. Viidennessä luvussa analysoidaan tutkimuksen tuloksia ja vastataan tutkimuskysymyksiin. Kuudennessa luvussa on tutkielman yhteenveto.

2 SUOSITTELUJÄRJESTELMÄT

Suosittelujärjestelmät ovat ohjelmallinen vastaus internetin alati kasvavan tietomäärän suodattamiseen. Suosittelujärjestelmällä tarkoitetaan ohjelmistoa, joka yrittää suodattaa tietoa erilaisin menetelmin, jotta se voisi suositella käyttäjälle sellaisia esineitä tai tuotteita, joita hän ei välttämättä olisi itse osannut löytää. Usein palveluiden valikoima on myös niin laaja, että siitä on vaikea saada käsitystä selaamalla tai etsimällä tiettyjä asioita. Suosittelujärjestelmä eroaa perinteisestä tiedonhakujärjestelmästä (kuten esimerkiksi hakukoneesta) siten, että sen tarkoitus on antaa käyttäjäkohtaisia suosituksia esimerkiksi pelkän asiasanahaun sijasta. Tätä kautta käyttäjä voi löytää dokumentteja, joita ei olisi muuten tullut löytäneeksi. Myös hakutuloksia voidaan järjestää siten, että käyttäjälle sopivimmat dokumentit näytetään tuloslistassa ensimmäisenä.

Kuvassa 2.1 on esitetty suosittelujärjestelmän toiminta periaatteellisella tasolla. Tarkemmin sanottuna kuvassa on esitetty *yhteisöllisen suosittelun* (engl. collaborative filtering) idea, joka on suosittu tapa suosittelujärjestelmien toteuttamiseen.



Kuva 2.1. *Yhteisöllisen suosittelujärjestelmän toiminta.*

Yhteisöllinen suosittelu perustuu käyttäjiltä saatuun arvostelutietoon, jonka perusteella suosittelujärjestelmä osaa arvottaa tuotteiden laatua ja osuvuutta käyttäjän omaan makuun. On olemassa muitakin tapoja, joista kerrotaan tarkemmin luvussa 2. Yhteistä kaikille suosittelujärjestelmille on kuitenkin se, että ne tutkivat jotain suositeltavan tuotteen piirrettä, jota verrataan käyttäjän mieltymykseen. Tämän jälkeen

järjestelmä voi tehdä suosituksen niistä tuotteista, joiden se arvelee tuovan käyttäjälle lisäarvoa. Yksinkertainen esimerkki suosittelusta on musiikkipalvelu, joka näyttää käyttäjän antaman artistin perusteella kappaleita samalta esittäjältä tai listan muista samankaltaisista artisteista. TV-tietokannasta voidaan yrittää löytää sellaisia sarjoja, joista käyttäjä olisi kiinnostunut. Tähän tietysti vaaditaan tietoa sekä käyttäjästä että etsittävän tiedon tyypistä.

Olennaista on, miten suosittelujärjestelmä saa tietoa käyttäjän tarpeista personoidun suosituksen tekemiseen. Suositus perustuu oletukseen tai päätelmään vastaanottajan tarpeista ja ominaisuuksista, joita käyttäjä on historiatiedon perusteella valikoinut (Sanastokeskus 2010). Sellaista yleismaailmallista suodatinta ei ole olemassa, joka soveltuisi kaikkien verkkopalvelujen käyttöön. Perusperiaatteena on yhdistää käyttäjän omat kiinnostuksen kohteet (esimerkiksi käyttäjän tekemien arvioiden tai selaushistorian perusteella) tietoon muun yhteisön kiinnostuksen kohteista. Näiden perusteella voidaan ennustaa sitä, miten käyttäjä arvioisi kohteita, joita hän ei vielä tunne. Perusoletus on se, että asiat joista käyttäjä aikaisemmin piti, ovat myös sellaisia asioita, joista käyttäjä tulee jatkossa pitämään (Su & Khoshgoftaar 2009).

Suosittelujärjestelmät jaetaan usein kahteen luokkaan sen perusteella, miten ne tietoa suodattavat. Suodatus voidaan tehdä joko tuotteen tietojen pohjalta (*sisältöpohjainen suodatus*, engl. content based filtering) tai muiden käyttäjien mielipiteen pohjalta, kuten kuvassa 2.1. Jotkin suosittelujärjestelmät toimivat myös näiden kahden yhdistelmänä. Suodatusmenetelmää, joka yhdistelee eri suodatustapoja kutsutaan *hybridisuodatukseksi* (engl. hybrid filtering) (Adomavicius & Tuzhilin 2005).

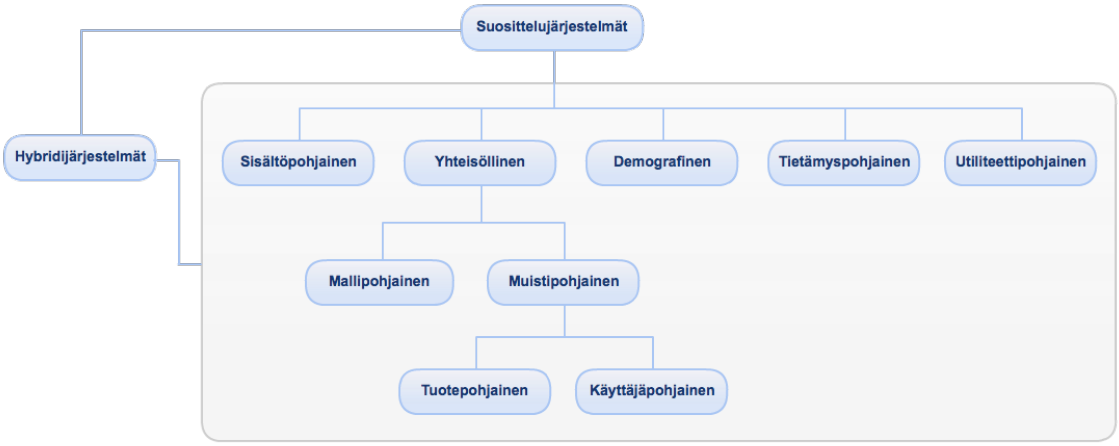
Tiedonhakujärjestelmissä käytetään yleensä sisältöön perustuvaa suodatusta, jossa käyttäjän haluamaa sisältöä haetaan avainsanojen perusteella (Nichols 1998). Sisällön perusteella tehtävä suosittelu vastaa luultavasti tarkemmin sitä, mitä käyttäjä haluaa, koska vertailu tehdään suoraan haetun tiedon sisältöön, jolloin vastaavuus on tarkempi. Tiedon keruuta ja analysointia on kuitenkin vaikeaa tehdä koneellisesti. Tämän vuoksi useimmissa suosittelujärjestelmissä käytetään muiden käyttäjien mielipiteeseen perustuvaa yhteisöllistä suositusta.

Yhteisöllistä suodatusta käyttävän järjestelmän ei tarvitse tietää mitään suositeltavasta kohteesta, vaan se suosittelee käyttäjälle sellaisia dokumentteja ja tuotteita, joista muutkin ovat pitäneet. Tyypillisesti suosittelujärjestelmässä on kaksi joukkoa – käyttäjät, jotka kaipaavat suosituksia sekä tuotteet, joita suositellaan (Kangas 2002). Suosittelu tapahtuu niin, että käyttäjät arvostelevat ensin joitain tuotteita. Arvosteluhistorian avulla voidaan ennustaa käyttäjälle arvio sellaisesta tuotteesta, jota hän ei ole vielä arvostellut.

Aiheesta on olemassa paljon tutkimuksia, koska suosittelujärjestelmien toteutukseen liittyy useita haasteita. Lisäksi reaalimaailman sovellutuksia on hyvin paljon. Henkilökohtaisia suositteluita toteutettaessa tuskin koskaan päästään täydellisiin tuloksiin. Parantamisen varaa kuitenkin on aina, ja tämä myös innostaa tutkijoita. Suosittelujärjestelmät olisi tulevaisuudessa hyvä myös saada toimimaan laajemmalla

alueella, sillä tällä hetkellä jokaisella palvelulla on oma, räätälöity tapansa tehdä suositteluja. (Adomavicius & Tuzhilin 2005)

Kuvassa 2.2 on selvennetty suosittelumenetelmien jakoa eri kategorioihin. Aliluvuissa 2.1-2.3 on kuvattu tarkemmin, mitä mikin kuvassa mainittu termi tarkoittaa.



Kuva 2.2. *Suosittelumenetelmien jaottelu.*

Suosittelujärjestelmät joutuvat täyttämään monenlaisia tarpeita. Käyttäjät haluavat erilaisia asioita, joihin vastaaminen on vaikeaa ja joskus jopa ristiriitaista. Tässä tutkielmassa keskitytään kuvaamaan käyttäjä- ja tuotepohjaisen suodatuksen toimintaa suosittelujärjestelmissä, suodatusmenetelmiin liittyviä ongelmia sekä näiden menetelmien yhdistämisellä saavutettavia hyötyjä. Tästä eteenpäin termi suosittelujärjestelmä viittakin järjestelmään, joka on toteutettu jonkin tässä dokumentissa esitellyn suodatusmenetelmän avulla.

2.1 Sisältöpohjainen suodatus

Sisältöpohjainen suosittelu perustuu dokumentista irrotettuun avainsana- tai piirretietoon dokumentin sisällöstä. Dokumenttia tai sen metatietoja analysoimalla voidaan tuottaa erilaisia luokittelusääntöjä, joiden perusteella dokumentin kiinnostavuutta voidaan ennustaa käyttäjän profiilin perusteella. (Kangas 2002) Tuotteen piirteiden esitystavasta riippuu, miten niitä saadaan yhdistettyä käyttäjäprofiileihin (Pazzani & Billsus 2007).

Perusperiaatteena on etsiä tuotteista tai dokumenteista piirteitä, joilla suositeltavia tuotteita voidaan kuvata. Esimerkiksi musiikkikappaleesta voidaan löytää tekijöiksi esittäjät, musiikin tyylilaji, tempo tai levy-yhtiö. Elokuvissa vastaavia ominaisuuksia ovat muun muassa näyttelijät, ohjaajat, tyylilajit ja aihepiirit. Suosittelun yhteydessä näitä piirteitä verrataan asioihin, joista käyttäjä on aikaisemminkin pitänyt. Mitä enemmän määreitä kutakin dokumenttia kuvaa, sitä tarkemmin ne voidaan kohdistaa tietylle käyttäjälle. Vastaavasti käyttäjän mieltymyksistä pitää olla varsin kattava kartta, jotta ennuste olisi käyttäjälle hyödyllinen. Tämänkaltainen piirteiden

erotus tuotekannasta on kuitenkin hyvin työlästä – etenkin, jos tuotteita ja käyttäjiä on hyvin paljon.

Sisältöpohjainen suosittelu nojaa pitkälti tiedonhakuun ja dokumentista saatavan tiedon louhintaan. Täten sopii parhaiten tekstisisältöihin kohdistuviin suosituksiin. Ero perinteisiin tiedonhakujärjestelmiin tulee suosittelussa käytettävien käyttäjäprofiilien myötä. Profiilit sisältävät tietoa käyttäjän mieltymyksistä, joita saadaan joko suoraan käyttäjältä itseltään kysymällä tai vaihtoehtoisesti käyttäjän toimia seuraamalla. (Adomavicius & Tuzhilin 2005)

Koska sisältöpohjaiset järjestelmät luottavat käyttäjän aikaisemmin tekemiin toimiin suosittelun tekemiseksi, kärsivät ne usein *alkuongelmasta* (engl. cold start). Tämä tarkoittaa sitä, että käyttäjä, jolla on vain vähän tallennettuja mieltymyksiä, ei voida saada kovinkaan tarkkoja suosituksia. Lisäksi tällaiset järjestelmät on suunniteltu suosittelemaan vain sellaisia asioita, jotka liittyvät selkeästi käsiteltävään dokumenttiin itseensä. Tämän vaatimuksen myötä sisällön tulisi olla helposti koneellisesti luettavissa – käytännössä siis tekstisisältöä. Jos tietoa dokumentin sisällöstä ei voida koneellisesti irrottaa (esimerkiksi kuvasta tai videotiedostosta), pitää tiedostoon liittyvät määreet ja metadata liittää siihen käsin. (Adomavicius & Tuzhilin 2005)

Avainsanoihin perustuvassa suosittelussa on myös ongelmana, ettei kahta samoin piirtein kuvattua dokumenttia voida erottaa toisistaan. Toisin sanoen hyvin kirjoitettu dokumentti ei eroa huonosta pelkän avainsana-analyysin avulla, jos ne käyttävät samoja termejä. (Adomavicius & Tuzhilin 2005)

2.2 Yhteisöllinen suodatus

Yhteisölliseen suodatukseen liittyvissä tutkimuksissa saatetaan aika ajoin käyttää myös termiä *sosiaalinen suodatus* (engl. social filtering) synonyymina tässä luvussa tarkemmin esitellylle yhteisölliselle suosittelulle. Vaikka näitä termejä käytetään usein ristiin, tarkoittaa termi sosiaalinen suosittelu tämän tutkielman puitteissa nimenomaan käyttäjien aktiivisesti toisilleen tekemää suosittelua sosiaalisten verkostojen välityksellä.

Ensimmäinen yhteisöllistä suodatusta käyttävä suosittelujärjestelmä oli Tapestry (Goldberg et al. 1992), josta menetelmän alkuperäinen nimi myös on peräisin. Nykyään yhteisöllistä suodatusta käytetään laajasti suosittelujärjestelmien ja verkkokauppojen toteutuksissa (Symeonidis et al. 2006). Suomenkielinen termi yhteisöllinen suodatus kuvaa menetelmää paremmin kuin alkuperäinen englanninkielinen termi, joka tarkoittaa yhteistoiminnallista suodatusta. Suositukset lasketaan käyttäjien antamista arvioista, mutta käyttäjät eivät silti tee yhteistyötä arvostelujen eteen. Usein käyttäjillä ei ole mitään suhdetta tai kontaktia toisiinsa vaan he toimivat kaikki omilla tahoillaan.

Toisin kuin sisältöpohjaisessa suosittelussa, yhteisöllisessä suodatuksessa käytetään suositusten laskemiseen muiden käyttäjien antamia arvioita. Arviot voidaan myös johtaa implisiittisesti heidän toimistaan ilman sen kummempaa mielenilmaisua käyttäjän puolelta. Yhteisöllisen suodatuksen hyvä puoli on se, että se ei ota kantaa suositeltavien tuotteiden sisältöön. Suositeltavana voi olla musiikkia, elokuvia,

palveluita tai vaikka muita käyttäjiä. Mitä tahansa webin sisältöä voidaan arvioida käyttäjien tekemien arvostelujen avulla.

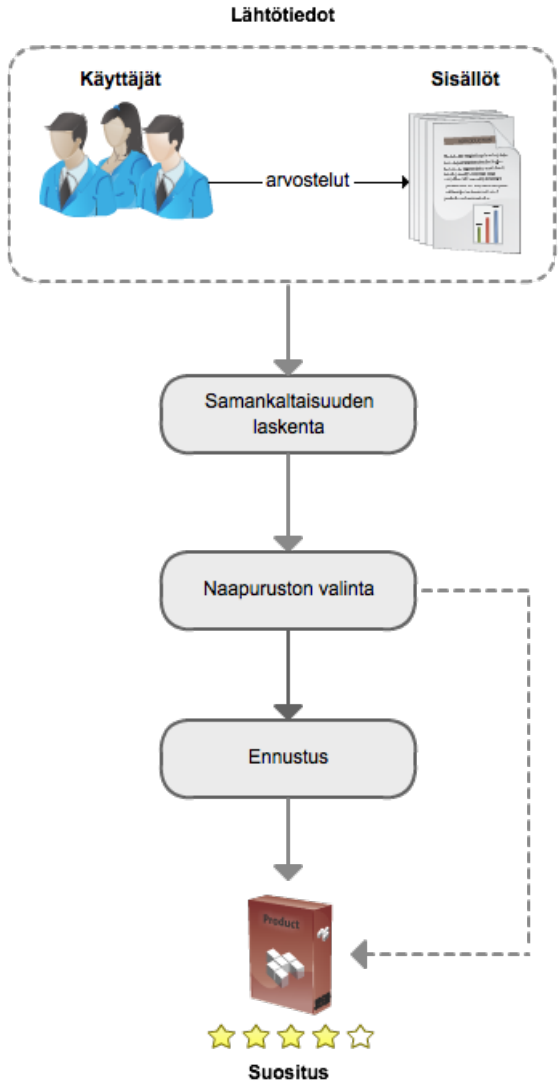
Vaikka käyttäjät eivät tuntisi toisiaan, voivat he silti olla samankaltaisia. *Samankaltaisuudella* (engl. similarity) tarkoitetaan suosittelujärjestelmien yhteydessä käyttäjien tekemien arvostelujen yhtenevyyttä eli korrelaatiota. Keskenään samankaltaisilla käyttäjillä todennäköisesti on myös hyvin samankaltainen maku arvosteltavien asioiden suhteen. Joka asiassa samankaltaisuus ei tietenkään päde – ihmisillä, jotka pitävät samankaltaisista elokuvista eivät välttämättä pidä samanlaisesta musiikista.

Yhteisöllinen suodatus voidaan jakaa kuvan 2.2 mukaisesti joko mallipohjaiseen tai muistipohjaiseen suodatukseen (Adomavicius & Tuzhilin 2005). Muistipohjainen suodatus jakaantuu edelleen käyttäjä- ja tuotepohjaiseen suodatukseen. Muistipohjaisessa suosituksessa käytetään hyväksi valmiiksi laskettuja tietoja käyttäjien tai tuotteiden välisistä suhteista, kun taas mallipohjaisessa menetelmässä suosittelumalli lasketaan ajonaikaisesti. Näitä menetelmiä yhdistelemällä voidaan yrittää parantaa suosittelujärjestelmän ennusteiden tarkkuutta (Wang et al. 2006). Kahta tai useampaa eri suosittelumenetelmää yhdistävää järjestelmää kutsutaan hybridisuosittelujärjestelmäksi.

Suosittelut perustuvat aikaisempiin tapahtumiin, arvosteluihin tai käyttäjiltä saatuun palautteeseen. Käyttäjä voi antaa arvion tuotteelle joko eksplisiittisesti tai implisiittisesti. Eksplisiittinen arvio tarkoittaa sitä, että käyttäjä kertoo suoraan, mitä pitää kustakin tuotteesta – esimerkiksi arvioimalla sen asteikolla yhdestä (”huono”) viiteen (”hyvä”). Arvion skaala on sovelluskohtainen valinta ja se voi olla yksi- tai kaksiarvoinen, tai moniarvoinen skalaari. Yksiarvoisessa eli unaarisessa arvostelussa arviolla on yksi arvo, esimerkiksi ”pidän tästä”. Kaksiarvoisessa eli binäärisessä skaalassa arvoja on kaksi, esimerkiksi ”pidän tästä” tai ”en pidä tästä”. Usein käytännön toteutuksissa arvoa merkitään numeroilla nolla (kielteinen) ja yksi (myönteinen). Moniarvoinen skalaari on jokin numero valitulta arvosteluväliltä.

Implisiittinen arvio perustuu käyttäjän käyttäytymisen ja hänen tekemisiensä tulkintaan. Esimerkiksi siirtyminen tuotesivuille voidaan katsoa kiinnostukseksi tuotetta kohtaan tai ostopäätös myönteiseksi arvioksi. Paljon sivulatauksia tai ostoja saanut tuote voidaan arvostaa korkeammalle kuin niitä vähän saanut. Kuten eksplisiittisessä arvioinnissa, skaala vaihtelee toteutuksen mukaan. Implisiittisen arvon ongelmana on sen tulkinta. Esimerkiksi ostopäätös ei kerro käyttäjän mausta, jos ostos on lahja jollekin toiselle.

Kuvassa 2.3 on esitetty yhteisöllisen suodatuksen eri vaiheet prosessina. Yhteisöllisen suodatuksen toiminta perustuu kolmeen asiaan – käyttäjiin, tuotteisiin sekä käyttäjien arvosteluihin näistä tuotteista. Nämä toimivat lähtötietoina, joiden perusteella voidaan laskea käyttäjien samankaltaisuus. Tämän vaiheen jälkeen voidaan jo tehdä suosituksia riippuen sovelluskohteesta. Joskus saattaa riittää, että suositellaan vain samankaltaisia tuotteita tai käyttäjiä. Yleensä kuitenkin laskettujen samankaltaisuuksien perusteella tehdään myös ennusteet sellaisille tuotteille, joita käyttäjä ei vielä tunne.



Kuva 2.3. *Yhteisöllisen suodatuksen prosessi.*

Kuvan 2.3 prosessimallissa samankaltaisuuden laskenta voidaan ajatella komponentiksi, jonka tilalle valitaan sovelluskohtaisesti siihen parhaiten sopiva laskentatapa. Samankaltaisuus voidaan siis laskea usealla eri tapaa, jonka jälkeen seuraavaan vaiheeseen valitaan käyttäjän kanssa eniten samankaltaiset naapurit. Samoin ennusteen laskemiseen on olemassa useita eri tapoja. Eri laskentamenetelmistä kerrotaan tarkemmin luvussa 2.2.1. Näiden kahden vaiheen algoritmien optimoinnilla voidaan yrittää parantaa järjestelmän suositusten määrällisesti mitattavaa laatua.

Järjestelmän tuotteet muodostavat joukon $I = \{i_1, i_2, i_3, \cdots, i_n\}$, jossa n on tuotteiden kokonaismäärä järjestelmässä, kun taas järjestelmän käyttäjät muodostavat joukon $U = \{u_1, u_2, u_3, \cdots, u_m\}$, jossa m on käyttäjien kokonaismäärä järjestelmässä. Käyttäjien tekemistä arvosteluista voidaan muodostaa arvostelumatriisi **R**, kuten taulukossa 2.1 on esitetty.

Taulukko 2.1. *Esimerkki arvostelumatriisista* ***R****.*

R	i_1	i_2	i_3	...	i_a	...	i_n
u_1	-	2	-		5		r_{u_1,i_n}
u_2	1	-	4		-		r_{u_2,i_n}
u_3	3	1	5		-		r_{u_3,i_n}
...							
u_a	-	2	4		1		r_{u_a,i_n}
...							
u_m	r_{u_m,i_1}	r_{u_m,i_2}	r_{u_m,i_3}		r_{u_m,i_a}		r_{u_m,i_n}

Taulukon 2.1 esimerkin arvot pohjautuvat perinteiseen, usein elokuvien arvostelussa käytettyyn viiden tähden arvosteluasteikkoon, jossa arvo $r = 1$ tarkoittaa huonoa ja arvo $r = 5$ hyvää arviota. Taulukon tyhjät kohdat tarkoittavat sitä, että käyttäjä ei ole vielä arvostellut kyseistä tuotetta (merkintä ”-”). Arvosteluskaalan valinta on sovelluskohtainen päätös, eikä mikään estä käyttämästä esimerkiksi asteikkoa $r = 1\ldots10$ tai binääristä asteikkoa $r \in [0,1]$. Tässä tapauksessa asteikoksi on valittu lukua 4 silmällä pitäen elokuvien arvostelussa usein käytetty viiden pisteen (tähden) asteikko, koska tutkielman aihe käsittelee elokuvien suosittelua.

Taulukon tyhjille kohdille voidaan myöhemmin ennustaa käyttäjän muiden arvostelujen perusteella arvio, jonka käyttäjä saattaisi antaa tuotteelle. Tätä ennustusarvoa voidaan käyttää suositusten tekemiseen hakemalla parhaat ennusteet saaneet tuotteet. Ennustamiseen tarvitaan tieto muista käyttäjän kanssa samankaltaisista käyttäjistä tai samankaltaisista tuotteista. Samankaltaisuus perustuu johonkin laskettuun etäisyys- tai korrelaatioarvoon. Riittää, että se on jokin numeerisesti ilmaistava arvo, jolla käyttäjät tai tuotteet voidaan järjestää.

Samankaltaisuus voidaan laskea arvostelujoukon perusteella, jota sisältää käyttäjien tekemät tai tuotteiden saamat arvostelut. Laskennan yhteydessä puhutaan usein aktiivisesta käyttäjästä tai tuotteesta. Tällä tarkoitetaan sitä käyttäjää tai tuotetta, jonka suhteen laskentaa ollaan kulloinkin tekemässä. Esimerkiksi aktiivisen käyttäjän u_a tekemät arvostelut (taulukon 2.1 korostettu rivi) muodostavat joukon $R_{u_a} = \left\{ r_{u_a,i_1}, r_{u_a,i_2}, r_{u_a,i_3}, \cdots, r_{u_a,i_n} \right\}$, jossa n on tuotteiden lukumäärä. Vastaavasti aktiivisen tuotteen i_a (taulukon 2.1 korostettu sarake) saamat arvostelut muodostavat joukon $R_{i_a} = \left\{ r_{u_1,i_a}, r_{u_2,i_a}, r_{u_3,i_a}, \cdots, r_{u_m,i_a} \right\}$, jossa m on käyttäjien lukumäärä.

Taulukossa 2.2 on esitetty esimerkki käyttäjien samankaltaisuusmatriisista S, jossa käyttäjien välille on laskettu heidän välisensä samankaltaisuusarvot välillä 0...1. Esimerkissä arvo 1 tarkoittaa täysin korreloivaa ja 0 ei lainkaan korreloivaa. Taulukon esimerkissä käyttäjät u_1 ja u_3 ovat arvostelujensa perusteella varsin samankaltaisia, sillä heidän korrelaatioarvonsa on varsin suuri (0.92). Vastaavasti käyttäjät u_1 ja u_2 ovat vähiten toistensa kaltaisia (korrelaatioarvo 0.14).

Taulukko 2.2. *Esimerkki käyttäjien samankaltaisuusmatriisista* $\mathbf{S}_u$.

$\mathbf{S}_u$	u_1	u_2	u_3	...	u_m
u_1	1	0.14	0.92		s_{u_1,u_m}
u_2	0.14	1	0.5		s_{u_2,u_m}
u_3	0.92	0.5	1		s_{u_3,u_m}
...					
u_m	s_{u_m,u_1}	s_{u_m,u_2}	s_{u_m,u_3}		s_{u_m,u_m}

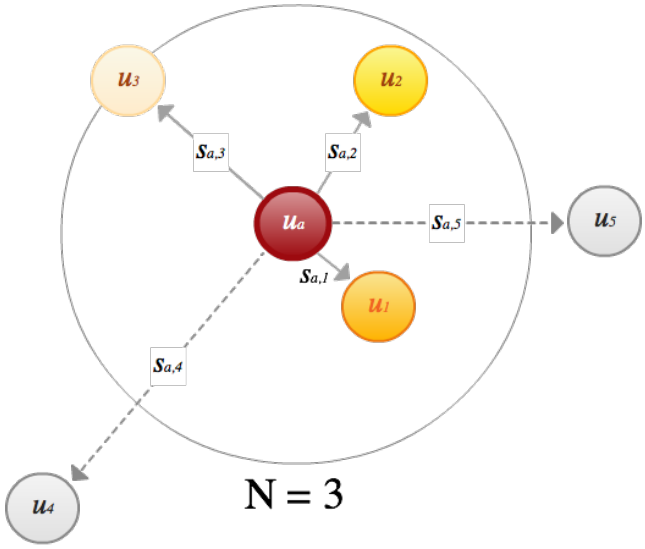
Taulukon 2.2 esimerkkiarvoja ei ole laskettu oikeista käyttäjistä vaan ne on esitetty vain esimerkin vuoksi. Samankaltaisuuden laskennassa lävistäjän arvoiksi tulee kuitenkin aina yksi. Tämä on luonnollista, sillä kukin käyttäjä on aina sataprosenttisesti samanlainen itsensä kanssa. Vastaavankaltainen matriisi voidaan muodostaa myös tuotteiden välille. Taulukossa 2.3 on esimerkki tällaisesta tuotteiden samankaltaisuusmatriisista. Kuten taulukossa 2.2, ovat taulukon 2.3 arvot myös keksittyjä.

Taulukko 2.3. *Esimerkki tuotteiden samankaltaisuusmatriisista* $\mathbf{S}_i$.

$\mathbf{S}_i$	i_1	i_2	i_3	...	i_n
i_1	1	0.25	0.07		s_{i_1,i_n}
i_2	0.25	1	0.43		s_{i_2,i_n}
i_3	0.07	0.43	1		s_{i_3,i_n}
...					
i_n	s_{i_n,i_1}	s_{i_n,i_2}	s_{i_n,i_3}		s_{i_n,i_n}

Käyttäjän u_a kanssa samankaltaiset käyttäjät muodostavat joukon $s_{u_a} = \left\{ s_{u_a,u_1}, s_{u_a,u_2}, s_{u_a,u_3}, \cdots, s_{u_a,u_m} \right\}$, jossa m on käyttäjien lukumäärä. Vastaavasti tuotteen i_a kanssa samankaltaiset tuotteet muodostavat joukon $S_{i_a} = \left\{ s_{i_a,i_1}, s_{i_a,i_2}, s_{i_a,i_3}, \cdots, s_{i_a,i_n} \right\}$, jossa n on tuotteiden määrä.

Ennusteen tekemisessä samankaltaisista käyttäjistä valitaan vain tietty määrä käyttäjiä. Tästä joukosta käytetään nimeä *naapurusto* (engl. neighbourhood). Naapuruston kokoa kuvataan arvolla N. Kuvassa 2.4 on esimerkki, jossa käyttäjän u_a kolmen eniten samankaltaisen käyttäjän naapurusto koostuu käyttäjistä u_1, u_2 ja u_3. Periaate on sama tuotteiden kanssa. Samankaltaisista tuotteista valitaan tietty määrä samankaltaisia tuotteita, joiden perusteella suosittelu tehdään.



Kuva 2.4. *Samankaltaisten käyttäjien valinta naapurustolla N = 3.*

Samankaltaisten käyttäjien joukkoa kutsutaan myös käyttäjän *lähimmiksi naapureiksi* (engl. nearest neighbour). Suosittelua, jossa ennustus tehdään *N*:n eniten samankaltaisen käyttäjän avulla, kutsutaan *top-N -suositteluksi* tai *lähimmän naapurin menetelmäksi* (Su & Khoshgoftaar 2009).

Laskettujen arvojen, eli samankaltaisten käyttäjien tai tuotteiden etsinnän ja ennusteen perusteella voidaan tehdä monenlaisia suosituksia. Käyttäjälle voidaan ehdottaa samankaltaisia käyttäjiä, jotka korreloivat hänen makunsa kanssa. Vastaavasti tuotteelle voidaan ehdottaa samankaltaisia tuotteita, joista muut käyttäjät ovat pitäneet. Varsinaisesta suosittelusta ja laskentamenetelmistä kerrotaan tarkemmin seuraavassa luvussa.

2.2.1 Muistipohjaiset suodatusmenetelmät

Muistipohjaiset suodatusmenetelmät käyttävät hyväkseen koko tuote- tai käyttäjäkantaa ennusteen laskennassa. Naapuruston samankaltaisuuden laskenta perustuu käyttäjien välisen etäisyyden tai korrelaation laskemiseen. Eräs yksinkertainen ja suosittu informaatiojärjestelmissä käytetty tapa on määritelmässä 2.1 esitetyn *kosinietäisyyteen* perustuvan samankaltaisuuden laskeminen (engl. cosine similarity). Matematiikassa sillä mitataan kahden vektorin välistä etäisyyttä ja sitä voidaan käyttää myös niiden välisen kulman laskemiseen.

Määritelmä 2.1. *Kosinietäisyys, jossa* $u \in U$ *ja* $v \in U$.

$$s_{u,v} = \frac{\sum_{i \in I} r_{u,i} r_{v,i}}{\sqrt{\sum_{i \in I} r_{u,i}^2} \sqrt{\sum_{i \in I} r_{v,i}^2}}$$

Määritelmässä 2.1 $r_{u,i}$ ja $r_{v,i}$ ovat käyttäjien u ja v arvostelut tuotteelle i. Kaavan osoittajassa tehdään pistetulo kahden vektorin välillä ja nimittäjässä vektorien pituudet kerrotaan keskenään. Saatu samankaltaisuus on arvo välillä s = 0…1, jossa arvo 1 tarkoittaa täysin samankaltaista ja 0 täysin erilaista. Tässä tapauksessa samankaltaisuus perustuu siis käyttäjien väliseen korrelaatioon, mutta kaava toimii myös tuotteiden samankaltaisuuden laskentaan.

Käyttäjien välinen arvosteluskaala vaihtelee usein eri henkilöiden välillä. Kosinietäisyyden laskennassa tämä voidaan ottaa huomioon vähentämällä käyttäjän tekemien arvostelujen keskiarvo käsiteltävästä arviosta. Tätä menetelmää kutsutaan *muokatuksi kosinietäisyydeksi* (engl. adjusted cosine similarity) ja sen kaava on esitetty määritelmässä 2.2.

Määritelmä 2.2. *Muokattu kosinietäisyys, jossa* $u \in U$, $v \in U$. *Termi* $\bar{r}$ *on käyttäjän tekemien arvostelujen keskiarvo.*

$$s_{u,v} = \frac{\sum_{i \in I}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i \in I}(r_{v,i} - \bar{r}_v)^2}}$$

Toinen suosittu menetelmä samankaltaisuuden laskentaan on *Pearsonin korrelaatio* (engl. Pearson correlation), jonka kaava on esitetty määritelmässä 2.3 (Su & Khoshgoftaar 2009). Se toimii käytännössä kuten muokattu kosinietäisyyskin. Erona on kuitenkin se, että käsiteltävien arvioiden joukko on rajattu vain niihin tuotteisiin, jotka molemmat ovat arvostelleet. Kaavan palauttamat arvot ovat välillä s = -1…1, jossa arvo 1 tarkoittaa täysin samankaltaista ja -1 täysin erilaista.

Määritelmä 2.3. *Pearsonin korrelaatio, jossa* $\hat{R} \in (R_u \cap R_v)$ *eli yhteisesti arvosteltujen tuotteiden joukko sekä* $u \in U$ *ja* $v \in U$.

$$s_{u,v} = \frac{\sum_{i \in \hat{R}}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in \hat{R}}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i \in \hat{R}}(r_{v,i} - \bar{r}_v)^2}}$$

Naapureiden etsinnän jälkeen voidaan erilaisia algoritmeja käyttämällä laskea ennuste sellaiselle tuotteelle, jota käyttäjä ei ole vielä arvostellut. Samankaltaisten käyttäjien arvosteluista voidaan myös poimia korkeimmalle arvostetut tuotteet ja muodostaa näistä N:n kappaleen suosituslista. Taulukossa 2.4 on esimerkki matriisista, jonka avulla ennusteet lasketaan.

Taulukko 2.4. *Esimerkki ennustematriisista* ***P****.*

P	i_1	i_2	i_3	...	i_a	...	i_n
u_1	?	2	?		5		p_{u_1,i_n}
u_2	1	?	4		?		p_{u_2,i_n}
u_3	3	1	5		?		p_{u_3,i_n}
...							
u_a	?	2	4		1		p_{u_a,i_n}
...							
u_m	p_{u_m,i_1}	p_{u_m,i_2}	p_{u_m,i_3}		p_{u_m,i_a}		p_{u_m,i_n}

Taulukossa 2.4 merkintä ”?” tarkoittaa arvioita tuotteelle, jota käyttäjä ei ole vielä arvostellut. Ennusteet lasketaan siis näille tuotteille. Numeroarvot ovat jo olemassa olevia arvioita tuotteille ja näitä käytetään hyväksi ennusteen laskemisessa. Naapuruston valinnan jälkeen voidaan ennustaa, mitä käyttäjä antaisi arvosanaksi tuotteelle, jota hän ei vielä tunne. Ennustus voidaan myös laskea monella tavalla. Määritelmässä 2.4 on esitetty yksi tapa ennusteen laskemiseen.

Määritelmä 2.4. *Ennusteen laskeminen. Kaavassa* $u_a \in U$ *on aktiivinen käyttäjä, jolle ennuste on tarkoitettu ja* $i_a \in I$ *tuote, jolle ennustus tehdään. Joukko* $\hat{S} = \left\{ s_{u_a,u_1}, \cdots, s_{u_a,u_N} \right\}$ *koostuu N:stä samankaltaisia käyttäjästä, eli valitusta naapurustosta.*

$$p_{u_a,i_a} = \frac{\sum_{u=\hat{S}} s_{u_a,u} r_{u,i_a}}{\sum_{u=\hat{S}} s_{u_a,u}}$$

Ennusteen laskennassa voidaan ottaa huomioon myös käyttäjän jo tekemät arvostelut, kuten määritelmässä 2.5 on esitetty.

Määritelmä 2.5. *Ennusteen laskeminen ottamalla huomioon aikaisempien arvostelujen keskiarvo. Kaavassa* $u_a \in U$ *on aktiivinen käyttäjä,* $i_a \in I$ *on tuote, jolle ennustus tehdään.* $\hat{S}$ *on joukko samankaltaisista käyttäjistä, jonka koko on N.* $\bar{r}$ *on käyttäjän tekemien arvostelujen keskiarvo.*

$$p_{u_a,i_a} = \bar{r}_{u_a} + \frac{\sum_{u=\hat{S}} s_{u_a,u} (r_{u,i_a} - \bar{r}_u)}{\sum_{u=\hat{S}} s_{u_a,u}}$$

Järjestämällä suositusten joukko $P_{u_a} = \left\{ p_{u_a,i_1}, \cdots, p_{u_a,u_k} \right\}$ (*k* on arvostelemattomien tuotteiden lukumäärä) voidaan valita suositeltaviksi parhaat ennusteet saaneet tuotteet.

2.2.2 Mallipohjaiset suodatusmenetelmät

Toisin kuin muistipohjaiset järjestelmät, mallipohjaiset menetelmät perustuvat tuote- tai käyttäjätietokannasta opittuihin malleihin. Rakennettuja malleja voidaan tämän jälkeen käyttää ennustuksien tekemiseen. Mallit käyttävät yleensä osaa aineistosta opetusaineistona mallin lähtötietoina. Mallia voidaan myös korjata tarvittaessa. (Adomavicius & Tuzhilin 2005)

Mallit perustuvat todennäköisyyksien ja odotusarvojen laskentaan. Mallin rakentamiseen voidaan käyttää erilaisia työkaluja, kuten esimerkiksi Bayesin verkkoja, *ryvästystä* (engl. clustering) tai sääntöpohjaista lähestymistapaa. Bayesin verkkomallissa käytetään hyväksi todennäköisyyksiä yhteisöllisessä suodatuksessa. Ryvästys on suosittu menetelmä, jossa samankaltaiset käyttäjät jaetaan samankaltaisuuden perusteella eri aliluokkiin (Helin & Koivisto 2010). Tarkoituksena on estimoida sitä todennäköisyyttä, että käyttäjä kuuluu johonkin luokkaan ja sen perusteella laskea arvioiden ehdollista todennäköisyyttä. Tavoitteena on siis ennustaa, mitä käyttäjä antaisi kullekin tuotteelle arvosanaksi ottaen huomioon hänen aikaisemman arvosteluhistoriansa.

Sääntöpohjaisissa menetelmissä etsitään assosiaatioita tuotteiden välillä, jotka useat käyttäjät ovat ostaneet. Tämän jälkeen tuotteita voidaan suositella assosiaation voimakkuuden perusteella.

2.3 Muita suosittelumenetelmiä

Tässä luvussa kerrotaan lyhyesti muista suosittelujärjestelmissä käytetyistä suodatusmenetelmistä. Tällaisia menetelmiä ovat muun muassa demografinen, utiliteettipohjainen, tietämyspohjainen suosittelu sekä hybridisuosittelu. Lista perustuu Viljasen (2006) kuvauksiin eri suosittelumenetelmistä.

Demografinen suosittelu perustuu käyttäjän ominaisuuksien mallintamiseen erilaisilla demografisilla mittareilla. Tällaisia ovat esimerkiksi ikä, sukupuoli tai varallisuus. Suosittelujärjestelmä pyrkii luokittelemaan käyttäjiä erilaisiin demograafisiin ryhmiin, joille suositellaan tietyntyyppisiä tuotteita. Esimerkiksi opiskelijaksi profiloituneelle käyttäjälle näytetään erilaisia tuotteita kuin urallaan jo edenneelle käyttäjälle. Tietoa demografisesta ryhmästä voidaan myös hyödyntää mainonnan kohdentamisessa.

Utiliteettipohjaisessa suosittelussa jokaiselle tuotteelle lasketaan hyödyllisyysarvo (utiliteetti), joka kuvaa tuotteen käyttäjälle tuomaa hyötyä. Tällainen hyötyominaisuus voi olla esimerkiksi tuotteen toimitusaika, jolloin nopeasti saatavilla olevia tuotteita voidaan suositella ensimmäiseksi käyttäjälle, joka haluaa tuotteensa mahdollisimman nopeasti. Edistyneessä käyttöliittymässä käyttäjä voi itse säätää, mitä ominaisuuksia haluaa painottaa omissa suositteluissaan.

Tietämyspohjaisessa suosittelussa yritetään mallintaa sitä, miten kukin tuote vastaa käyttäjän tarpeita. Esimerkiksi ravintoloita suositellessa niitä voidaan kuvata ominaisuuksilla kuten tyyli (”hiljainen”) tai sijainti (”merinäköala”).

Hybridisuodatusta käyttävissä suosittelujärjestelmissä yhdistellään eri sisältöpohjaisen ja yhteisöllisen suosittelun metodeja. Tarkoituksena on yhdistää kahden tai useamman järjestelmän hyvät puolet ja siten kompensoida toisen menetelmän huonoja puolia. Esimerkiksi yhteisöllinen suodatus yhdistetään usein jonkin toisen suodatusmallin kanssa, jotta alkuongelmilta vältyttäisiin.

Tässä tutkielmassa käytetään yhteisölliseen suodatukseen perustuvaa muistipohjaista menetelmää, jonka antamia suosituksia pyritään parantamaan tiedolla sosiaalisesta verkostosta. Kyseessä on siis eräänlainen hybridisuosittelujärjestelmä. Tässä luvussa mainittuja menetelmiä ei käsitellä enää tämän tarkemmin tutkielman puitteissa.

2.4 Suosittelujärjestelmien ongelmat ja haasteet

Koneellinen suosittelu on vaikeaa ja eri menetelmissä on omat haasteensa. Näiden ongelmien ratkaisemiseen on kehitetty useita ratkaisuja ja tutkimuksia on olemassa runsaasti. Tässä luvussa esitellään joitain tyypillisiä, suosittelujärjestelmien toteutukseen liittyviä ongelmia.

Harvuuden ongelma (engl. sparsity) on yleinen ongelma monelle suosittelujärjestelmälle – etenkin sellaiselle, jossa on suuri valikoima erilaisia tuotteita. Käytännössä tämä tarkoittaa sitä, että edes hyvin aktiivisella käyttäjällä on todennäköisesti arvosteltuna (tai ostettuna) melko pieni prosenttiosuus koko tuotemäärästä (Gong 2010). Tämän takia saattaa olla vaikeata löytää käyttäjälle samankaltaisia naapureita, joiden perusteella suosittelu tapahtuu, eikä tarkkoja suosituksia siis voida tehdä. Usein onkin hyödyllistä yhdistää yhteisölliseen suodatukseen muita suositteluteknikoita, kuten sisällöllistä suodatusta.

Alkuongelmalla (engl. cold start) tarkoitetaan tilannetta, jossa järjestelmään tulee uusi tuote tai käyttäjä. Näihin liittyviä arvioita ei vielä luonnollisesti ole, joten samankaltaisten käyttäjien tai tuotteiden naapurustoa ei voida laskea (Gong 2010). Uuden käyttäjän on mahdotonta saada tarkkoja suosituksia, jos hän ei ole arvostellut yhtään tuotetta. Vastaavasti ilman arvosteluja oleva tuote hautautuu helposti massan sekaan. Alkuongelma pätee uusien tuotteiden lisäksi myös käyttäjille, jonka maku on hyvin vaihteleva tai ailahteleva. Yksi ratkaisu on pyytää uutta käyttäjää arvostelemaan joukko tuotteita rekisteröitymisen yhteydessä tai muuten kuvaamaan makuaan. Tuotteiden kohdalla järjestelmä voi tuoda esiin tuotteita, joissa on vähän tai ei yhtään arvostelua ja pyytää käyttäjää arvostelemaan ne.

Skaalautuvuus (engl. scaling) tulee vastaan järjestelmissä, joissa on laaja määrä tuotteita ja käyttäjiä. Lähimmän naapurin menetelmää käyttävät algoritmit hyödyntävät dataa koko laajuudessaan, joten tuote- ja käyttäjäkantojen kasvaessa kasvaa myös laskennan vaatima aika ja teho. Naapureiden laskennasta muodostuukin yleensä

muistipohjaisten menetelmien pullonkaula. Siksi useimmat järjestelmät kannattaa toteuttaa niin, että todellinen, suurimman tehon vaativa laskenta tapahtuu jossain järjestelmän taustalla (Adomavicius & Tuzhilin 2005). Näin ollen suosituksia tehdessä voidaan käyttää valmiiksi laskettuja arvoja.

Poikkeukselliset yksilöt ovat käyttäjiä, joiden maut saattavat olla hyvinkin poikkeavat järjestelmissä havaituista ryhmistä, joten he eivät hyödy yhteisöllisen suodatuksen eduista. Tällaisia käyttäjiä kutsutaan harmaiksi lampaiksi (engl. gray sheep). Mustat lampaat (engl. black sheep) taas ovat käyttäjiä, joiden maku on niin idiosynkraattinen eli muusta käyttäjämassasta poikkeava, että ennustaminen on mahdotonta. Vaikka ongelma onkin järjestelmän, on mustien lampaiden olemassaolo järjestelmässä yleensä hyväksyttävä virhe. (Su & Khoshgoftaar 2009)

Suosituimmuussuuntauma (engl. popularity bias) tarkoittaa sitä, että järjestelmä osaa suosittaa vain tuotteita, jotka ovat suosittuja. Tämä taas johtuu siitä, että tällaisilla tuotteilla on paljon (hyviä) arvosteluja. Jos käyttäjän maku on pitkälti yleisen mielipiteen mukainen, voi olla vaikea tehdä suosituksia käyttäjälle muista tuotteista. Käyttäjä jää tällöin vaille niin kutsuttuja onnellisia löytöjä, joita käyttäjät yleensä haluavat suositusjärjestelmältä.

Hakkerointi (engl. hacking, shilling attacks) on mahdollista järjestelmissä, joissa arvioita voi tehdä kuka vain. Arvostelemalla omia tuotteita korkeilla arvosanoilla ja kilpailijoiden tuotteita huonoilla voidaan yrittää vaikuttaa suosittelujärjestelmän antamiin suosituksiin. Eräs tapa vaikuttaa yhden käyttäjän suosituksiin on profiilin kopiointi. Jos käyttäjät ovat arvostelleet samat tuotteet samoilla arvosanoilla, ovat nämä käyttäjät järjestelmän mielestä täysin samankaltaisia. Tämän jälkeen hakkeri voi arvostella tuotteita, joita haluaa mainostaa. Koska käyttäjät ovat järjestelmän mielestä samankaltaisia, saavat kopioprofiilin arvostelut suuremman painon suosittelussa.

Suurin osa yllä mainituista asioista liittyy käyttäjän havaitsemaan tai kokemaan arvioon suositusten laadusta. Tätä asiaa käsitellään tarkemmin seuraavassa luvussa.

2.5 Suosittelujärjestelmän arviointi

Suosittelujärjestelmien toteutuksessa on tärkeää asettaa mittareita, joiden avulla suositusten tehokkuutta tarkkaillaan. Mitattu laatu voi olla joko määrällistä eli kvantitatiivista tai laadullista eli kvalitatiivista tietoa. Määrällinen tieto perustuu laskelmiin järjestelmän antamista ennusteista jollekin tietylle käyttäjälle, joita sitten verrataan käyttäjän antamiin oikeisiin arvoihin. Laadullinen arviointi perustuu käyttäjän kokemaan arvoon suositusten laadusta.

2.5.1 Määrällinen arviointi

Suosittelujärjestelmän tehokkuutta voidaan arvioida numeerisesti monin eri tavoin. Yleinen tapa informaationhakujärjestelmissä on käyttää osaa olemassa olevista arvioista opetusaineistona ja loppuosaa järjestelmän arvioimiseen. Esimerkiksi 60% aineistosta voidaan valita käytettäväksi laskentaan, jolloin loppuosaa käytetään ennustuksen

vertaamiseen oikeaan tulokseen. Suosittuja mittareita ovat muun muassa määritelmässä 2.6 esitetty *absoluuttinen keskimääräinen virhe* (MAE, engl. mean absolute error) ja määritelmässä 2.7 esitetty *keskineliövirheen neliöjuuri* (RMSE, engl. root mean square error).

Määritelmä 2.6. *Absoluuttinen keskimääräinen virhe (MAE), jossa joukko* $P = \left\{ p_{u_a,i_1}, \cdots, p_{u_a,i_n}, \right\}$ *koostuu käyttäjälle tehdyistä ennusteista, k on ennusteiden lukumäärä ja* $r \in R_{u_a}$.

$$MAE_{u_a} = \frac{\sum_{j \in P} \left| p_{u_a,i_j} - r_{u_a,i_j} \right|}{k}$$

Määritelmä 2.7. *Keskineliövirheen neliöjuuri (RMSE), jossa* $p_{u,i}$ *on ennuste käyttäjän u arvostelulle tuotteesta i ja* $r_{u,i}$ *on käyttäjän u antama oikea arvio tuotteelle i.*

$$RMSE_u = \sqrt{\frac{\sum_{i \in P} (p_{u,i} - r_{u,i})^2}{n}}$$

Edellä mainitut määrälliset mittarit ovat usein käytössä informaationhakujärjestelmissä ja ne perustuvat piirretietoon (yhteisöllisen suodatuksen tapauksessa arvostelumatriisiin) ja siitä laskettujen arvojen vertailuun. Suosittelujärjestelmissä, jotka perustuvat lähimmän naapurin menetelmiin on samankaltaisten käyttäjien löytäminen järjestelmän kriittisin vaihe.

Arvostelumatriisista lasketun korrelaation arvoa laskettaessa on otettava huomioon, että kyse on vain arvosanojen korreloinnista. Tämän perusteella ei voida suoraan sanoa käyttäjien makujen tai persoonien korreloivan vaikka arvosanoista laskettu korrelaatioarvo olisikin suuri. Tämän vuoksi yhteisöllisen suodatuksen järjestelmät ovat haavoittuvaisia monille luvussa 2.4 esitetyille ongelmille, esimerkiksi profiilin kopioinnille.

Käytetty arvosteluskaala vaikuttaa omalta osaltaan korrelaation laskentaan. Jos käytössä on jokin muu kuin yksikäsitteinen arvosteluasteikko, tulee vastaan käyttäjien erilainen käsitys arvosteluperiaatteista. Esimerkiksi arvosteluasteikolla 1…10 joku saattaa hyödyntää koko skaalaa kun toinen taas käyttää vain kouluarvosanoja (4…10) arvosteluun. Yhden käyttäjän mielipide voi myös vaihdella ajan tai tunteiden vaihteluiden myötä. Ongelma on tärkeä osa järjestelmän suunnittelua ja kalibrointia. Käyttäjien samankaltaisuuden laskennassa käyttäjien välistä skaalaa voidaan yrittää normalisoida, kuten esimerkiksi määritelmissä 2.2 ja 2.3 on esitetty.

Cosley et al. (2003) havaitsivat tutkimuksessaan, että käyttäjät arvostelevat tuotteita yleensä samoin – valitusta arvosteluskaalasta riippumatta. Samassa tutkimuksessa he kuitenkin havaitsivat, että usein käyttäjät säätävät arvioitaan ennustettua arvoa kohti, jos tuotteen voi arvostella suosituksen yhteydessä. Tällainen käyttäjän manipulointi saattaa olla tahatonta, mutta vaikuttaa negatiivisesti koko

suositusjärjestelmän kykyyn tehdä tarkkoja suosituksia. Järjestelmä, joka ohjaa käyttäjän antamaan ennustetta vastaavan arvosanan tuotteelle saa todennäköisesti parempia arvosanoja virhemetriikkojen perusteella. Todellinen suositusten laatu saattaa kuitenkin peittyä hyvien virhetulosten alle.

2.5.2 Laadullinen arviointi

Käyttäjän kokema suositusten laatu on tärkeä, joskaan ei kovin helposti mitattavissa oleva suosittelujärjestelmän arviointiperuste. Käyttäjät tarvitsevat luotettavia suosituksia, jotta he löytävät haluamansa tuotteet. Suositusjärjestelmä, joka ei pysty toistuvasti antamaan hyviä suosituksia on käytännössä hyödytön. (Sarwar et al. 2001) Siksi onkin tärkeää, että käyttäjä saa tietoonsa suosittelun perusteet. Käyttäjä voi mahdollisesti myös antaa palautetta järjestelmän antamista suosituksista.

Swearingenin ja Sinhan (2006) mukaan suosittelujärjestelmä tuo käyttäjälleen arvoa, jos se täyttää alla olevat vaatimukset.

1. Valaa luottamusta
2. Tarjoaa yksityiskohtaisia tietoja suosittelun syistä
3. Suosittelujärjestelmän toimintalogiikka on (ainakin osittain) läpinäkyvää
4. Laajentaa käyttäjän horisonttia esittelemällä uusia tuotteita
5. Tarjoaa mahdollisuuden säätää suosituksessa käytettyjä parametreja (poistaa joitain tuotteita tai lisätä niitä)

Laadullinen tiedon mittaaminen suosittelujärjestelmissä on vaikeampi toteuttaa. Joitain arvoja on mahdollista päätellä implisiittisesti – esimerkiksi ostopäätökseen johtanut suosittelu on merkki suosittelun onnistumisesta. Tarkka palaute vaatii kuitenkin jonkinlaista käyttäjäkyselyä tai mahdollisuutta antaa palautetta järjestelmän toiminnasta. Takaisinkytkentä palautteen antamiseen vaatii sekä työpanosta palvelun kehittäjiltä sekä aktiivista toimintaa käyttäjältä. Tämä taas on osin ristiriidassa automaattisesti toimivan järjestelmän kanssa.

On olemassa monia huomioon otettavia seikkoja, joita ei voida algoritmillisesti mitata. Schafer et al. (2007) mainitsevat näistä muun muassa uutuusarvon, kattavuuden, oppimiskäyrän sekä uskottavuuden.

Uutuusarvo (engl. novelty) kuvaa käyttäjän kokemaa hyötyä uusien tuotteiden suosittelusta. Suositus voi olla käyttäjän maun mukainen, mutta yleensä käyttäjä haluaa saada suosituksia asioista, jotka eivät ole niin tuttuja ja tuovat täten käyttäjälle uutuusarvoa. Odottamattomat suositukset eli ns. *onnelliset löydöt* (engl. serendipitous discovery) ovat yksi yhteisöllisen suodatuksen hyvistä puolista. Nämä ovat tuotteita, joita käyttäjä ei ole aikaisemmin kohdannut tai edes tullut harkinneeksi. (Schafer et al. 2007) Uutuusarvoon pyrkivät suosittelujärjestelmät vaativat kuitenkin paljon tietoa käyttäjästä sekä toimivan palautejärjestelmän, jolla käyttäjä voi indikoida suosituksen osuvuuden. Tämä voi tarkoittaa esimerkiksi nappia suosittelun yhteydessä, jonka avulla käyttäjä voi ilmaista mielipiteensä siitä, oliko suositus hyödyllinen vai ei.

Kattavuus (engl. coverage) tarkoittaa sitä prosenttiosuutta tuotteista, joille suosittelujärjestelmä voi tehdä arvioennustuksia. Käyttäjien tai heidän tekemiensä arvioiden harvuuden perusteella kaikille tuotteille ei voida ennustaa luotettavaa arviota.

Oppimiskäyrä (engl. learning rate) mittaa sitä, kuinka nopeasti suositusjärjestelmä pystyy tarjoamaan luotettavia suosittelutuloksia käyttäjän maun mukaisesti. Yleensä tämä tarkoittaa arvostelujen määrää, joka tarvitaan ennen kuin laatu nousee tarvittavalle tasolle.

Uskottavuus (engl. confidence) mittaa suosittelujärjestelmän kykyä mitata antamiensa suositusten laatua. Suurin osa yhteisöllistä suodatusta käyttävistä järjestelmistä pohjaa suosituksensa suurimman todennäköisyyden saaneeseen ennustukseen. Järjestelmä, joka voi mitata omaa uskottavuuttaan voi rajoittaa antamiaan suosittelujaan vain sellaisiin tuloksiin, joiden se uskoo vastaavan käyttäjän omaa arviota. Tämä johtaa suurempaan osumatarkkuuteen, mutta samalla kattavuus ja uutuusarvo saattaa kärsiä.

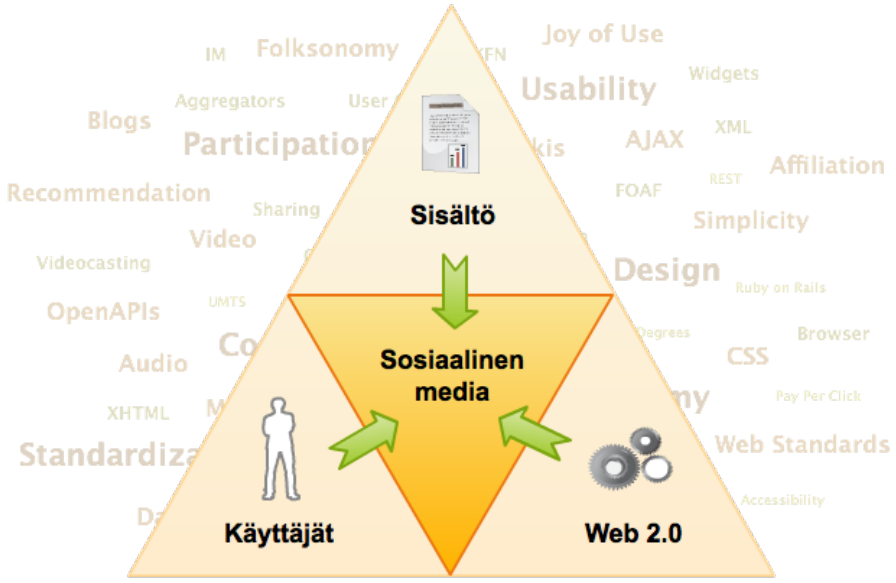
Sinha ja Swearingen (2001) ovat havainneet, että käyttäjät kokevat usein ystäviensä suosittelut arvokkaammaksi kuin automaattisen suosittelujärjestelmien. Yksi keino lisätä luottamusta järjestelmään onkin tarjota käyttäjälle tietoa muista lähipiirin käyttäjistä, jotka ovat suositelleet tuotetta. Sosiaalista verkostoa ja sen hyödyntämistä suosittelussa tarkastellaan tarkemmin seuraavissa luvuissa.

3 SOSIAALINEN MEDIA JA SOSIAALISET VERKOSTOT

Sosiaalinen media (engl. social media) on internetin teknologioita ja palveluita hyödyntävä viestinnän muoto, jossa pääosassa ovat käyttäjälähtöinen vuorovaikutus ja sisällöntuotanto. Tärkeitä asioita ovat myös käyttäjien väliset suhteet ja niiden luonti. (Sanastokeskus 2010) Tieto- ja viestintäteknologian nopea kehitys on tuonut mukanaan monia uusia tapoja ihmisten väliseen kommunikointiin ja verkostoitumiseen. Koska sosiaalinen media on hyvin laaja käsite, keskitytään tässä tutkielmassa lähinnä vain sellaisiin palveluihin, joissa on mahdollista luoda verkostoja käyttäjien välille, ja näiden verkostojen hyödyntämiseen verkkopalvelun suositteluominaisuuksiin.

Sosiaalisen median tunnuspiirteitä ovat vuorovaikutteisuus, käyttäjälähtöisyys, teknologiasidonnaisuus, yhteisten merkitysten rakentaminen sekä vaikutukset yhteiskuntaan, talouteen ja kulttuuriin (Erkkola 2008). Lietsalan ja Sirkkusen (2008) tavoitteena on ollut muodostaa käsitys siitä, millaisia *lajityyppejä* (engl. genre) sosiaalisen median palveluissa on tunnistettavissa verrattuna perinteiseen mediaan. Eri lajityyppiin kuuluvat sosiaalisen median palvelut eroavat toisistaan sen suhteen, miten ne motivoivat käyttäjiä, millaisia toimintamahdollisuuksia ne tarjoavat käyttäjilleen sekä millaisessa roolissa käyttäjä on. Yhteistä sosiaalisen median palveluille on kuitenkin se, että ne pyrkivät edistämään *käyttäjäsisältöjen* (engl. user-generated content) julkaisua ja jakelua.

Sosiaalisesta mediasta ja sen tuomista mahdollisuuksista käytetään usein myös käsitettä *web 2.0*. Tällä viitataan uusien verkkoteknologioiden mahdollistamiin palveluihin, kuten blogeihin, käyttäjien itsensä luomaan sisältöön ja sen jakamiseen, maksuttomuuteen ja yhteisöllisyyteen (Hintikka 2007). Käytännössä termi web 2.0 on teknologinen viitekehys, jonka puitteissa käyttäjät voivat tuottaa sisältöjä ja muodostaa erilaisia yhteisöjä (Kangas 2007). *Sosiaaliset ohjelmistot* (engl. social software) puolestaan ovat web 2.0 -sovelluksia, joiden avulla luodaan sosiaalisia verkostoja ja hallitaan niiden sisältöjä (Lietsala & Sirkkunen 2008) Kuvassa 3.1 on eräs määritelmä sosiaalisen median muodostumisesta, joka kuvaa edellä mainittujen käsitteiden nivoutumista (Kangas et al. 2007).



Kuva 3.1. *Sosiaalisen median elementit (Kangas et al. 2007).*

Tässä tutkielmassa ei kuitenkaan oteta kantaa termien eroon. Tutkielman puitteissa sosiaalisilla verkostoitumispalveluilla ja -sivustoilla tarkoitetaan sellaisia palveluita, joiden tarkoituksena on kerätä joukko käyttäjiä ja sen jälkeen rakentaa kytköksiä näiden käyttäjien välille. Kytkökset voivat ilmentää käyttäjien suhteita todellisessa maailmassa tai ne voivat olla siitä kokonaan irrallisia. Käyttäjien välille syntyy suhteita heidän kiinnostuksensa, mielipiteidensä tai muiden aktiviteettiensa kautta.

Yhteistä kaikille tämäntapaisille verkkopalveluille on käyttäjien profiilitiedon tallennus. Profiili voi sisältää palvelusta riippuen monia asioita, mutta yleensä sen kautta käyttäjä hallitsee kiinnostuksensa kohteita ja verkoston suhteita. Sosiaalisissa verkostoissa olennaista on myös mahdollisuus jakaa ideoita, aktiviteetteja tai tapahtumia ja kommentoida toisten kiinnostuksen kohteita.

Sosiaalisen median suosion kasvun myötä pinnalle ovat nousseet *sosiaaliset verkostoitumispalvelut* (engl. social networking sites, SNS*)*. Näissä palveluissa pääpaino on käyttäjien välisellä interaktiolla. Palvelut kuten Facebook[3], Twitter[4], MySpace[5] ja LinkedIn[6] ovat tuoneet itsensä osaksi miljoonien ihmisten päivärutiinia. Katsauksessaan sosiaalisiin verkostoitumispalveluihin boyd ja Ellison (2007) määrittävät verkostoitumispalvelun palveluksi, joka mahdollistaa oman (ainakin osittain julkisen) profiilin luonnin, kontaktien luonnin muihin käyttäjiin listojen kautta sekä omien ja muiden käyttäjien kontaktilistan selailun.

Kuten sosiaalinen mediakin, on sana ”verkostoitumispalvelu” myös eräänlainen sateenvarjokäsite, joka kattaa monta erilaista toteutusta ja palvelua. Tarkka luokittelu on

[3] http://www.facebook.com
[4] http://www.twitter.com
[5] http://www.myspace.com
[6] http://www.linkedin.com

kuitenkin vaikeaa, ja ilmiö saattaa olla jo liian iso omaksi yläkäsitteeksensä. Toisaalta voidaan myös puhua vain ominaisuudesta, joka on liitetty palvelun yhteyteen. Sosiaalinen verkosto saattaa siis olla vain sosiaalisen median tai web 2.0 -termin alle kuuluva käsite, kuten wikit tai folksonomiat, eikä oma yläkäsitteensä. (Tienvieri 2010)

3.1 Sosiaaliset verkostoitumispalvelut

Sosiaalisten verkostoitumispalvelujen aihe on usein rajattu johonkin tiettyyn aihepiiriin, kuten esimerkiksi musiikkiin (Last.fm[7]), yritysverkostoitumiseen (Linkedin) tai kirjoihin (Goodreads[8]). Tunnetuin näistä lienee Facebook, joka on poikkeava siinä mielessä, että se ei ole keskittynyt mihinkään tiettyyn kiinnostusalueeseen, vaan ainoastaan sosiaaliseen verkostoitumiseen. Facebook-palvelusta on nykyään tullut lähes synonyymi sosiaaliselle verkostoitumispalvelulle – käyttäähän sitä jo yli 800 miljoonaa ihmistä (Facebook 2012a). Facebook on penetroitunut myös valtamedioihin ja mainoksissa näkee usein palvelun logon verkkosivun osoitteen vierellä. Facebookin levinneisyyttä on epäilemättä auttanut myös sen tarjoama erittäin monipuolinen rajapinta, jonka avulla muut palvelut voivat käyttää hyväksi Facebookin dataa.

Sosiaalisissa verkostoitumispalveluissa korostuvat ihmisten väliset suhteet, kommunikaatio ja yhteysverkostojen luonti. *Heikot siteet* (engl. weak ties) käyttäjien välillä ja avoin ympäristö ovat auttaneet palveluiden suosion kasvussa. Heikoilla siteillä tarkoitetaan yhteyksiä ihmisten välillä, joilla on erilaisia sosiaalisia piirteitä, kuten uskomuksia tai kiinnostuksia. (Wever et al. 2007). Sosiaalisessa mediassa tosimaailman fyysiset rajoitukset eivät päde, joten ainakin teoriassa ihmisellä on koko maailma saavutettavissa työpöydältänsä käsin. Nykyaikaisessa sosiaalisessa verkkopalvelussa verkostoituminen on helppoa ja vaivatonta. Tämä voi myös tapahtua huomaamattomasti vuorovaikutuksen myötä uusien ihmisten kanssa.

Käyttäjät voivat verkostoitua oma-aloitteisesti lisäämällä toisia käyttäjiä omaan sosiaaliseen verkostoonsa tai seuraamalla heidän tilapäivityksiään. Suhteet käyttäjien välillä voivat olla joko yhden- tai kahdensuuntaisia (Lee & Brusilovsky 2009). Tällä tarkoitetaan sitä, että toinen käyttäjä ”seuraa takaisin” käyttäjää, joka lisäsi tämän. Monissa palveluissa käyttäjän lisäys sosiaaliseen verkostoon on automaattisesti kahdensuuntainen toimenpide, sillä yhteyttä ei muodosteta ennen kuin toinen antaa siihen luvan. Toisen käyttäjän voi siis halutessaan torjua tai tehdä näkyvyysrajoitteita omalle profiilisivulle.

Käyttäjän sosiaaliset verkostot voivat olla joko käyttäjän itsensä määrittämiä tai sellainen voidaan implisiittisesti johtaa esimerkiksi keskustelualueen viestitiheyksistä (Marttila 2010). Yuan et al. (2009) käyttävät tutkimuksessaan termiä *eksplisiittinen sosiaalinen suhde* (engl. explicit social relationship) kuvaamaan verkostoa, joka on käyttäjän itsensä rakentama. Toisin sanoen käyttäjä aktiivisesti itse lisää uusia käyttäjiä verkostoonsa. Tässä tutkielmassa keskitytäänkin vain sellaisiin palveluihin, joissa

[7] http://www.last.fm
[8] http://www.goodreads.com/

verkosto on käyttäjän eksplisiittisesti määrittämä. Golbeck (2007) asettaa *web-pohjaiselle sosiaaliselle verkostolle* (engl. web-based social networ*k*, WBSN) neljä tunnusmerkkiä:

1. Palvelua käytetään web-selaimella
2. Käyttäjät määrittävät sosiaaliset suhteet eksplisiittisesti
3. Järjestelmä tukee eksplisiittisten kontaktien luontia
4. Käyttäjien väliset suhteet ovat näkyvillä ja selattavissa

Suurin osa suosituimmista verkostoitumispalveluista täyttävät nämä määreet, kuten esimerkiksi Facebook, LinkedIn tai Last.fm. Tämän tutkielman kannalta merkityksellisiä ovat sellaiset palvelut, jotka tarjoavat ohjelmallisen pääsyn verkostoon palvelun ulkopuolelta. Seuraavissa aliluvuissa on listattu joitain sellaisia palveluita, jotka täyttävät edellä mainitut kriteerit. Tärkeimmäksi näistä on tässä nostettu Facebook, sillä se on sosiaalisista verkostoitumispalveluista selvästi suosituin ja se tarjoaa kattavan rajapinnan verkoston hyödyntämiseen. Muitakin palveluita on olemassa, mutta niitä ei tässä tutkielmassa käsitellä.

3.1.1 Facebook

Facebook on suosittu sosiaalinen yhteisöpalvelu ja sillä on maailmanlaajuisesti jo noin 800 miljoonaa käyttäjää (Facebook 2012a). Käyttäjä rekisteröityy palveluun omalla nimellään ja luo oman profiilin. Profiili sisältää yleensä kuvan käyttäjästä sekä syntymäpäivän, siviilisäädyn ja koulutushistorian tapaista tietoa. Tämän jälkeen käyttäjä voi ottaa yhteyttä muihin käyttäjiin ja lisätä heidät omaan ystävälistaansa. Myös muut käyttäjät voivat ottaa yhteyttä. Ystäviksi tuleminen vaatii molempien osapuolien hyväksynnän.

3.1.2 Google+

Google+ on vuonna 2011 julkaistu Googlen kehittämä yhteisöpalvelu (Gundotra 2011). Palvelu on samantapainen kuin Facebook ja se keskittyykin ihmisten välisiin, jokapäiväisiin sosiaalisiin suhteisiin. Toisin kuin Facebookissa, on Google+:ssa mahdollisuus jakaa ystäviä ja kontakteja erilaisiin *piireihin* (engl. circles). Piirien kautta käyttäjä voi jakaa tiedon kätevästi tarkalleen valitun ihmisryhmän kesken.

3.1.3 Last.fm

Last.fm on musiikin suosittelupalvelu, jossa on mukana myös sosiaalisia ominaisuuksia (Last.fm 2012). Palvelun ideana on auttaa käyttäjää löytämään uutta musiikkia sen perusteella, mitä hän jo kuuntelee. Tiedot voidaan joko syöttää käsin tai sivustolta ladattavan ohjelman avulla. Ohjelma rekisteröi tiedot jokaisesta kuunnellusta kappaleesta sivustolle automaattisesti. Sivustolla on myös paljon sosiaalisia

ominaisuuksia, ja käyttäjä voi yhdistää itsensä muihin käyttäjiin, artisteihin tai yhteisöihin.

3.1.4 LinkedIn

LinkedIn on yritysmaailman verkostoitumiseen keskittyvä palvelu, joka aloitti toimintansa vuonna 2003 (LinkedIn 2012). Kyseessä on maailman suurin ammattilaisverkosto, jolla on yli 120 miljoonaa käyttäjää yli 200 maassa. Käyttäjän palveluun luoma profiili toimii käytännössä verkkopohjaisena ansioluettelona. Profiiliin kuuluu olennaisesti työ- ja koulutushistoria, taidot ja harrastukset sekä kiinnostuksen kohteet. Käyttäjän lisääminen omaan kontaktilistaan vaatii molempien tahojen hyväksynnän. Sivustolla on useita ammatillisia erityisryhmiä, joihin käyttäjä voi liittyä.

3.2 Sosiaalisten verkostojen teoriaa

Ihmisen sosiaalinen verkosto muodostuu hänen suhteistaan toisiin ihmisiin. Vastaavasti hänen tuttavillaan on omat verkostonsa, jotka voivat olla osin päällekkäisiä. Tunnetun teorian mukaan jokainen ihminen maapallolla on enintään kuuden askeleen (ihmisen) päässä toisistaan (Milgram 1967). Sosiaalisella verkostolla voidaan kuvata mitä vain suhteita, kuten esimerkiksi ystävyyssuhteita tai vaikka työpaikan hierarkkisia suhteita.

Verkosto koostuu joukosta toimijoita eli *solmuista* (engl. node, vertex), jotka yhdistyvät toisiinsa jonkin *suhteen* (engl. relation) kautta. Verkostoanalyysissa solmut voivat kuvata mitä vain, kuten esimerkiksi ihmisiä, yrityksiä tai valtioita. (Kankainen & Salminen 2011) Solmujen väliset suhteet riippuvat kontekstista ja voivat kuvata mitä tahansa käyttäjien välillä tapahtuvaa suhdetta. Yhteys voi perustua esimerkiksi ystävyyteen tai sukulaisuussuhteeseen, vuorovaikutuksen määrään tai ryhmän jäsenyyteen. Suhteen luonne voi myös riippua näkökulmasta, eli se voi olla yksi- tai kaksisuuntainen. Myös suhteen voimakkuus voi vaihdella näkökulmasta riippuen. Käyttäjä, jota toinen pitää luotettavana ei välttämättä luota häneen yhtä paljon.

Ihmisille tyypillisesti verkostoja voi olla lukuisia riippuen kontekstista, verkon toimijoista ja heidän välisistään suhteista. Tässä tutkielmassa keskitytään kuitenkin vain ihmisten välisiin suhteisiin verkkopalveluissa. Tällaista verkostoa kutsutaan *sosiaaliseksi verkostoksi* (engl. social network). Sosiaalisia verkostoja voidaan tulkita monin keinoin, kuten esimerkiksi matemaattisin menetelmin tai graafivisualisoinnin avulla. *Sosiaalisten verkostojen analyysi* (engl. social network analysis, SNA) on sosiaalisten verkostojen tutkintaan keskittynyt tutkimussuunta, joka pyrkii analysoimaan näitä verkostoja sekä yksittäisten toimijoiden että koko verkon tasolla. SNA:n analyysimenetelmät pohjautuvatkin vahvasti sekä graafiteoriaan että matriisilaskentaan. Graafiteoriaa käytetään enemmän verkoston mallintamiseen ja visualisointiin, kun taas matriisilaskentaa hyödynnetään enemmän verkostojen laskennallisessa analyysissa. (Wasserman & Faust 1994)

Matriisien avulla on helppo esittää laajakin verkosto koneellisessa muodossa ja tehdä sille laskutoimituksia. Sosiaalisesta verkostosta muodostetussa matriisissa vaaka-

ja pystyakselilla ovat verkon toimijat. Risteyskohdassa on numero, joka kuvaa toimijoiden välisen suhteen voimakkuutta. Tällaista matriisia kutsutaan *sosiomatriisiksi* (engl. sociomatrix). Yhteys voi olla joko kaksiarvoinen tai arvotettu. Kaksiarvoisessa yhteydessä arvo 1 merkitsee yhteyttä ja 0 tarkoittaa, että yhteyttä ei ole. Arvotetussa yhteydessä lukuarvolla voidaan kuvata tarkemmin suhteen voimakkuutta.

Taulukossa 3.1 on esitetty esimerkit sekä kaksiarvoisesta että arvotetusta sosiomatriisista. Taulukon arvot ovat keksittyjä ja esitetty vain esimerkin vuoksi. Taulukon arvot on esitetty asteikolla 0…1, mutta tämäkin on sovelluskohtainen ratkaisu. Todellisuudessa voidaan käyttää mitä tahansa asteikkoa, joka parhaiten kuvaa verkon rakennetta ja vastaa sovelluksen tarpeita.

	u_1	u_2	u_3	u_4
u_1	-	0	1	1
u_2	1	-	0	1
u_3	1	0	-	0
u_4	0	1	0	-

(a)

	u_1	u_2	u_3	u_4
u_1	-	0.4	0	1
u_2	0.7	-	0.5	0.5
u_3	1	0.25	-	0.75
u_4	0.1	0	1	-

(b)

Taulukko 3.1. *Kaksiarvoinen (a) ja arvotettu (b) sosiomatriisi.*

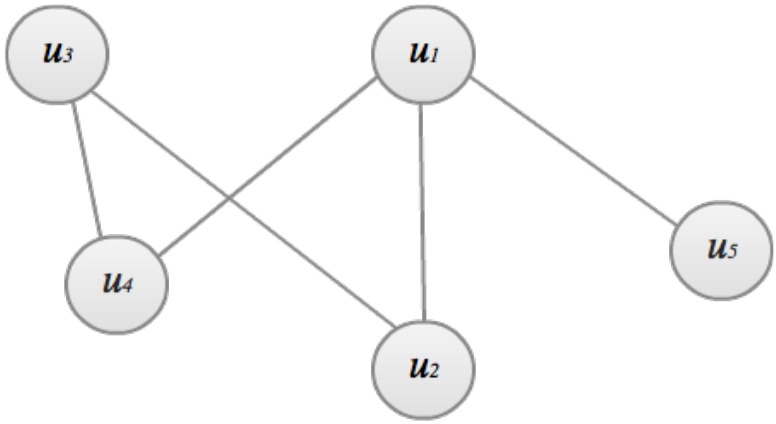
Kuten taulukosta 3.1 on nähdään, yhteydet toimijoiden välillä eivät välttämättä ole kaksisuuntaisia. Käyttäjällä u_1 on yhteys käyttäjään u_2, mutta ei toisinpäin. Sosiaalisissa verkostoissa yhteydet kuitenkin ovat usein kaksisuuntaisia ja yhteyden luominen kahden käyttäjän välillä vaatii molempien hyväksynnän. Tällöin matriisin arvot ovat symmetrisiä eri toimijoiden välillä.

Yleisin käytetty tekniikka verkostojen visualisoinnissa on graafi. Graafien avulla on helppo visualisoida verkoston rakennetta ja sisäisiä suhteita. Graafit koostuvat solmuista ja kaarista (tai nuolista). Solmut kuvaavat verkoston toimijoita, jotka ovat sosiaalisen verkoston tapauksessa ihmisiä. Kaaret kuvaavat vuorostaan toimijoiden välisiä suhteita. Graafiteorian menetelmin on myös helppo tehdä laskutoimituksia verkoston erilaisten reittien laskemiseen, esimerkiksi lyhyin etäisyys solmusta toiseen.

Graafi voi olla suunnattu tai suuntaamaton. Suunta kuvaa solmujen assosiaation suuntaa, jos suhde on yksipuolinen. Esimerkiksi solmu A vaikuttaa solmu B:hen, mutta solmu B ei voi vaikuttaa solmu A:han. Graafien kuvaamisessa voidaan myös hyödyntää erityyppisiä tai -paksuisia viivoja assosiaatioiden vahvuuksien kuvaamiseen. Tälläisiä graafeja kutsutaan multigraafeiksi. (Freeman 2009)

Graafi voi olla yksi- tai kaksimoodinen (tai enemmänkin) riippuen siitä, kuinka montaa tyyppiä verkoston solmut voivat esittää. Yksimoodinen verkosto voi esimerkiksi koostua vain ihmisistä, kun taas kaksimoodisessa verkostossa toimijoina voivat olla ihmiset sekä tapahtumat. Tässä tutkielmassa keskitytään kuitenkin vain yksimoodisten verkostojen käsittelyyn, joissa toimijoina ovat ihmiset, eli suosittelujärjestelmän

käyttäjät. Tällaisista verkostoista käytetään nimitystä *sosiogrammi* (engl. sociogram). Kuvassa 3.2 on esimerkki suuntaamattoman verkoston sosiogrammista.



Kuva 3.2. *Suuntaamattoman verkoston sosiogrammi.*

Kuvan 3.2 esimerkissä toimija u_1 on yhteydessä toimijoihin u_2, u_4 ja u_5, mutta ei toimijaan u_3. Käyttäjä u_2 on yhteydessä toimijoihin u_1 ja u_3. Toimija u_3 on yhteydessä toimijoihin u_2 ja u_4. Vastaavasti toimija u_4 on yhteydessä toimijoihin u_1 ja u_3. Toimija u_5 on yhteydessä vain toimijaan u_1. Tässä esitetyt yhteydet toimivat siis molempiin suuntiin, koska graafi on suuntaamaton.

Tässä tutkielmassa käsiteltävät verkostot ovat suunnattuja siinä mielessä, että vuorovaikutus käyttäjien välillä tapahtuu aina molempiin suuntiin. Käyttäjä A ei voi olla käyttäjän B kontakti ilman B:n lupaa. Yhteyden luomisen jälkeen A:sta tulee automaattisesti myös B:n kontakti. Jossain verkostoissa (esimerkiksi Twitterissä) käyttäjä voi lisätä toisen käyttäjän kontaktikseen, eli hän *seuraa* (engl. following) tätä käyttäjää. Tällöin kyseessä on yhdensuuntainen yhteys. Nämä ovat kuitenkin rajattu tutkielman aiheen ulkopuolelle. Käytännössä käsiteltävät verkostot ovat suuntaamattomia, sillä oletuksella että yhteys toimii molempiin suuntiin.

Sosiaalinen graafi on eräänlainen kartta henkilöistä, jotka käyttäjä tuntee. Tunnettavuuden käsite saattaa vaihdella suurestikin riippuen verkoston kontekstista. Esimerkiksi sosiaalinen graafi voi kertoa perheenjäsenten tai työyhteisön suhteista. Tällaisia verkostoja syntyy lähinnä sosiaalisissa palveluissa kuten Facebook ja LinkedIn, jotka keskittyvät sosiaalisten siteiden hallintaan.

Kiinnostusgraafi kertoo siitä, mistä asioista käyttäjä pitää. Esimerkiksi Last.fm-palvelusta voidaan hakea tiedot artisteista, joista käyttäjä pitää. Perinteisesti nämä kaksi asiaa ovat olleet erillään. Verkkokaupat pitävät kirjaa siitä, mitä asioita käyttäjä on ostanut tai mistä hän on pitänyt. Sosiaaliset verkostoitumispalvelut taas kartoittavat käyttäjän sosiaalisten kontaktien verkostoa. Se, että käyttäjä saa tietoa asioista, joita käyttäjän kontaktit pitävät, ei välttämättä takaa, että käyttäjä itse pitää samoista asioista. Tulevaisuuden web-sovelluksissa käyttäjien kiinnostusgraafit ja sosiaalinen graafi tulevat yhdistymään yhdeksi ja tätä kautta saadaan aikaan myös parempia suosituksia.

3.3 Liittymät sosiaalisiin verkostoitumispalveluihin

Kolmannen osapuolen kehittäjän näkökulmasta nykyisten sosiaalisten verkostojen hyvä puoli on se, että palveluiden sisältämää dataa on ryhdytty tarjoamaan käytettäväksi myös palvelun ulkopuolella. Jotkut palvelut tarjoavat ohjelmallisen pääsyn verkkodataan suoraan rajapintakutsujen kautta tai aineistot saa ladattua pienellä vaivalla koneellisesti luettavaan muotoon. Kaikki luvussa 3.1 mainitut palvelut tarjoavat erillisen rajapinnan kehittäjien käyttöön, jolla palvelun tietoja voidaan hakea tai jopa muokata palvelun ulkopuolelta.

Perustavanlaatuinen ongelma sosiaalisten verkostojen hyödyntämisessä on se, että ei ole olemassa vain yhtä palvelua tai rajapintaa verkkojen käsittelyyn. Erilaisia palveluja on lukuisia ja ne eivät ole useinkaan missään yhteydessä toisiinsa. Jotkut palvelut, kuten Facebook tai LinkedIn, mahdollistavat kontaktien etsinnän sähköpostiosoitteiden avulla tai olemassa olevan verkoston tuonnin palveluun. Verkostot eivät kuitenkaan pysy synkronoituna ilman käyttäjän tekemää manuaalista työtä.

Verkostot eivät myöskään ole tasavertaisia. Brad Fitzpatrick (2007) on esittänyt ajatuksen *solmujen tasavertaisuudesta* (engl. node equivalence). Tämä mahdollistaisi sen, että solmujen tunnisteet pysyisivät samana verkostojen välillä. Tällä tavoin esimerkiksi Facebook-verkostosta voisi poimia ne kontaktit, jotka ovat myös työkavereita, täydennettynä LinkedIn-verkostosta löydetyillä kollegoilla. Solmujen arvot ovat siis samoja kahden eri palvelun välillä, mutta ainoastaan konteksti, eli yhteys solmujen välillä vaihtuu. Samalla olisi helpompaa uuteen palveluun liittyessä etsiä, ketkä kontakteista ovat jo palvelussa.

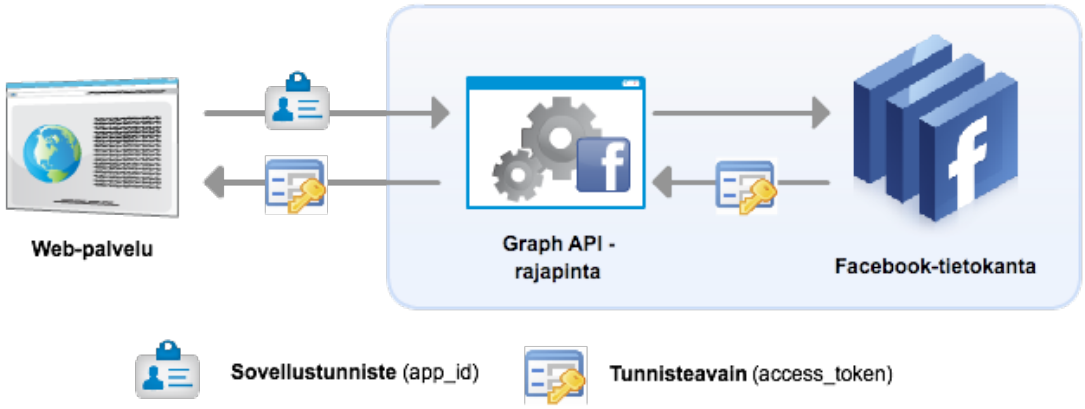
Koska yhdenmukaista rajapintaa tai mahdollisuutta yhdistää sosiaaliset verkostot ei vielä ole, on tässä tutkielmassa keskitytty Facebookin tarjoamiin palveluihin. Facebook tarjoaa kattavan valikoiman ohjelmointirajapintoja, joiden avulla palvelun ulkopuolelta voidaan käsitellä käyttäjien tai muiden sivujen tietoja. Facebookin kirjautumistietoja voidaan hyödyntää liittymien avulla muissa palveluissa. Facebook toimii myös ulkopuolisten sovellusten ajoalustana. Ulkoisia sovelluksia voidaan ajaa Facebookin ”sisällä”, eli niitä ajetaan kehysikkunassa, mutta ne näyttävät olevan osa Facebook-sivustoa.

Facebookin tarjoamat liittymät jakautuvat kolmeen kategoriaan: sosiaalisiin liitännäisiin, sovelluksiin ja Graph API -rajapintaan. *Sosiaaliset liitännäiset* (engl. social plugins) ovat Facebookin tarjoamia valmiita koodinpalasia, joita voidaan liittää mille tahansa verkkosivulle. Tällaisia liitännäisiä ovat muun muassa tykkää-nappi (engl. like button), kommenttiosio (engl. comments), kavereiden tapahtumasyöte (engl. activity feed) tai kirjautumislomake (engl. login button). (Facebook 2012b) Nämä ovat valmiita työkaluja, joita ei tarvitse itse toteuttaa. Riittää, että kehittäjä liittää saamansa koodin omalle verkkosivustolle. Kaikki tarpeellinen tieto ja toteutuslogiikka on toteutettu Facebookin puolesta.

Sovellukset (engl. applications) on tarkoitettu ajettavaksi Facebookin tarjoaman ympäristön sisällä. Sovellus ajetaan Facebookin näkymässä, mutta oman kehyksensä sisällä, johon haetaan sisältö ulkoiselta palvelimelta. Vaikka sovellusta ajetaankin Facebookin ulkopuolella, tarjoaa se mahdollisuuden integroitua Facebookin eri osa-alueisiin, kuten käyttäjien tietoihin tai heidän toimiinsa Graph API -rajapinnan kautta.

Graph API -rajapinta tarjoaa pääsyn kaikkeen Facebookin sisältämään sosiaaliseen tietoon. Rajapinta perustuu sosiaaliseen graafiin, jossa jokaisella toimijalla on oma tunnisteensa. Rajapinta on suunniteltu käytettäväksi joko suoraan HTTP-pyyntöjen avulla tai JavaScript-kirjastoja käyttäen. Tunnistautumaton käyttäjä pääsee käsiksi vain julkiseen tietoon, mutta kirjautumalla saa näkyviin yksityisempää informaatiota. Tosin kirjautunutkin käyttäjä pääsee käsiksi vain siihen tietoon, mikä hänelle olisi normaalistikin käytettävissä Facebookin käyttöliittymän kautta. Rajapinta toimii siis aina yksittäisen käyttäjän kontekstissa, ja käyttöoikeudet määräytyvät sen mukaan.

Kuvassa 3.3. on havainnollistettu rajapinnan käyttöä. Web-palvelu voi käyttää rajapintaa Facebookin myöntämän sovellustunnuksen avulla. Tämän jälkeen palvelun toteuttaja pääsee käsiksi tietokannan tietoihin rajapintakutsujen kautta. Kutsuille on kuitenkin välitettävä tunnisteavain, jolla rajataan tiedonhakua vain aktiiviseen käyttäjään.



Kuva 3.3. *Facebookin Graph API -rajapinnan käyttö (Tamada 2011).*

Facebookin tarjoama rajapinta mahdollistaa luvussa 3.2 mainitun sosiaalisen graafin käytön erilaisissa sovelluksissa. Rajapinnan tarjoama monimoodinen sosiaalinen graafi sisältää kaikki käyttäjää koskevat tiedot, kuten esimerkiksi hänen kiinnostuksen kohteensa ja ryhmät, joihin hän kuuluu. Yhteyksiä eri solmujen välillä voidaan tarkastella yhdessä ja erikseen. Myös muiden käyttäjien julkiset tiedot ovat saatavilla.

3.4 Sosiaalinen suosittelu

Sosiaalisen suosittelun käsite on yksinkertainen, onhan se kaikille tuttu asia tosielämästä. Ihmiset suosittelevat toisilleen jatkuvasti hyväksi kokemiaan asioita joko oma-aloitteisesti tai vastauksena toisen tiedusteluun. Esimerkiksi Heikki voi suositella Kaijalle katsomaansa elokuvaa, jonka Kaija katselee myöhemmin Heikin suosituksesta. Opettaja voi suositella oppilailleen hyvää oppikirjaa. Tällaisessa tapauksessa paljon merkitystä on sillä, mikä on suosittelijan ja suositusta kaipaavan henkilön suhde. Sosiaalisessa suosittelussa tämä *henkilöiden välinen vaikutus* (engl. interpersonal influence) onkin kriittisessä asemassa. Esimerkiksi uutta mekkoa ostava tyttö saattaa jättää mekon ostamatta, jos suosittelijana on hänen äitinsä. Toisaalta, jos hänen poikaystävänsä sanoo pitävänsä siitä, saattaa hän nyt ollakin eri mieltä mekon ostamisesta. (Huang et al. 2010)

Sosiaalisten verkostojen hyödyntäminen automaattisten suosittelujen tekemiseen on varsin uusi asia, eikä tutkimuksia ole saatavilla kovinkaan paljon. Sosiaalisten verkostoitumispalveluiden suosion kasvun myötä on kuitenkin helppo nähdä miksi tarve tähän on kasvanut. Koska sosiaaliset verkostot, kuten Facebook tai Google+, mallintavat tosielämän ystävyyssuhteita, on luonnollista olettaa, että käyttäjä haluaa saada suosituksia omilta ystäviltä myös verkkosivustoilla. Erikoistuneemmissa palveluissa, kuten Last.fm-palvelussa, yhteisöt kasvavat yhteisten kiinnostuksien ympärille.

Tavanomaisessa, yhteisöllisen suodatuksen avulla toteutetussa suosittelujärjestelmässä haetaan käyttäjien tekemien arvioiden perusteella korreloivia käyttäjiä. Samankaltaisten käyttäjien tekemien arvioiden avulla voidaan suositella aktiiviselle käyttäjälle entuudestaan tuntemattomia tuotteita. Nämä järjestelmät kärsivät kuitenkin luvussa 2.5 mainituista ongelmista, kuten datan harvuudesta tai alkuongelmasta. Tutkimukset (esimerkiksi Huang 2010) ovat osoittaneet, että käyttäjän omia sosiaalisia verkostoja hyödyntämällä voidaan parantaa suosittelutuloksia. Vaikka kaikkia ongelmia ei verkostojen avulla pystytäkään kiertämään – ne saattavat myös lisätä niitä – voidaan niiden avulla kuitenkin parantaa perinteisen yhteisöllisen suodatuksen tulosta ympäristöissä, joissa tietoa sosiaalista suhteista on saatavilla.

Joissain järjestelmissä, kuten Golbeckin (2006) esittelemässä FilmTrust-järjestelmässä, käyttäjä voi määrittää kuinka paljon kuhunkin käyttäjään luottaa. Luottamus määritellään asteikolla 1-10, jossa 1 on pienin ja 10 suurin mahdollinen luottamusta kuvaava luku. Luottamus tarkoittaa tässä tapauksessa sitoutumista toimintaan sillä uskomuksella, että toisen käyttäjän toimet johtavat hyvään tulokseen. Esimerkiksi Kaija luottaa Heikkiin elokuvien suhteen, jos Kaija katsoo elokuvan Heikin suosituksen pohjalta.

Konstan ja Riedl (2003) ovat jakaneet yhteisöllistä suodatusta hyödyntävien järjestelmien toiminnan kolmeen eri kategoriaan: aktiiviseen hakuun, aktiiviseen suositteluun sekä automaattiseen suositteluun. *Aktiivisen haun* (engl. pull-active)

järjestelmissä käyttäjän rooli aktiivisena toimijana suositusten saamisessa korostuu. Käyttäjä hakee itse järjestelmästä tietoa tai tekee suosituskyselyitä. Yhteisön jäsenellä on myös mahdollisuus tehdä omia arvioitaan tiedoista tai tuotteista.

Yhtymäkohta nykyajan sosiaalisiin palveluihin on esimerkiksi Twitter, jossa käyttäjä voi määrittää viesteihinsä (engl. tweet) kanavia (engl. hashtag), joissa viestit näkyvät. Kanava merkitään ristikkomerkillä ”#”. Kanavan nimeä klikkaamalla käyttäjä pääsee lukemaan kaikki kyseisellä merkkijonolla merkityt viestit. Kanavasta muodostuu tällöin aktiivinen hakutermi. Twitterissä myös käyttäjien seuraus (engl. following) on eräänlaista sosiaalista hakua, sillä käyttäjä etsii ja valitsee itse henkilöt, joilta haluaa saada suosituksia, tai tässä tapauksessa mielenkiintoisia viestejä. (Wei et al. 2011)

Aktiivisessa suosittelussa (engl. push-active) järjestelmän käyttäjät voivat tietoja selatessaan lähettää suosituksia mielenkiintoisista aiheista muille käyttäjille. Toimintaa voidaan verrata sähköpostiketjuihin, joissa ystävät lähettelevät toisilleen hauskoja kuvia tai mielenkiintoisia artikkeleita. Nykyiset sosiaaliset verkostoitumispalvelut mahdollistavat hyvin tämänkaltaisen toiminnan. Esimerkiksi Facebookissa käyttäjä voi kirjoittaa toisen käyttäjän seinälle (engl. wall post), lähettää tälle viestejä tai jakaa sisältöä omalla seinällään. Facebookin tunnus, peukalon kuvan sisältävä tykkää-painike on varmasti tuttu useille verkon käyttäjille. Tykkäämällä jostakin verkkosivusta tai artikkelista voidaan kertoa muille kavereilleen hyvästä asiasta.

Automaattisissa (engl. automated) yhteisöllisen suosittelun järjestelmissä käytetään hyväksi tietokantaa, johon käyttäjien arvostelut on tallennettu. Näiden perusteella voidaan etsiä samankaltaisia käyttäjiä ja tehdä suositteluja heidän mielipiteensä perusteella, kuten luvussa 2 on kuvattu. Suurin osa verkkopalveluissa käytettävistä suosittelujärjestelmistä ovat automaattisia. Etuna on se, että suosittelut toimivat myös anonyymeissä yhteisöissä. Aktiivinen haku ja suosittelu toimivat pienissä yhteisöissä, joiden jäsenillä on samankaltaiset kiinnostuksen kohteet. Yleensä nämä käyttäjät myös tuntevat toisensa ja tietävät, mistä he olisivat kiinnostuneita.

Sosiaalisen suosittelun tapoja on monia erilaisia, joista tässä luvussa on esitelty muutamia. Jotkut näistä ovat varsin yksinkertaisia, kuten esimerkiksi viestin lähettäminen toiselle käyttäjälle sosiaalisen verkostoitumispalvelun kautta. Tutkielman varsinainen tarkoitus on kuitenkin keskittyä koneellisesti tehtävään, automaattiseen suositteluun sosiaalisia piirteitä hyödyntäen.

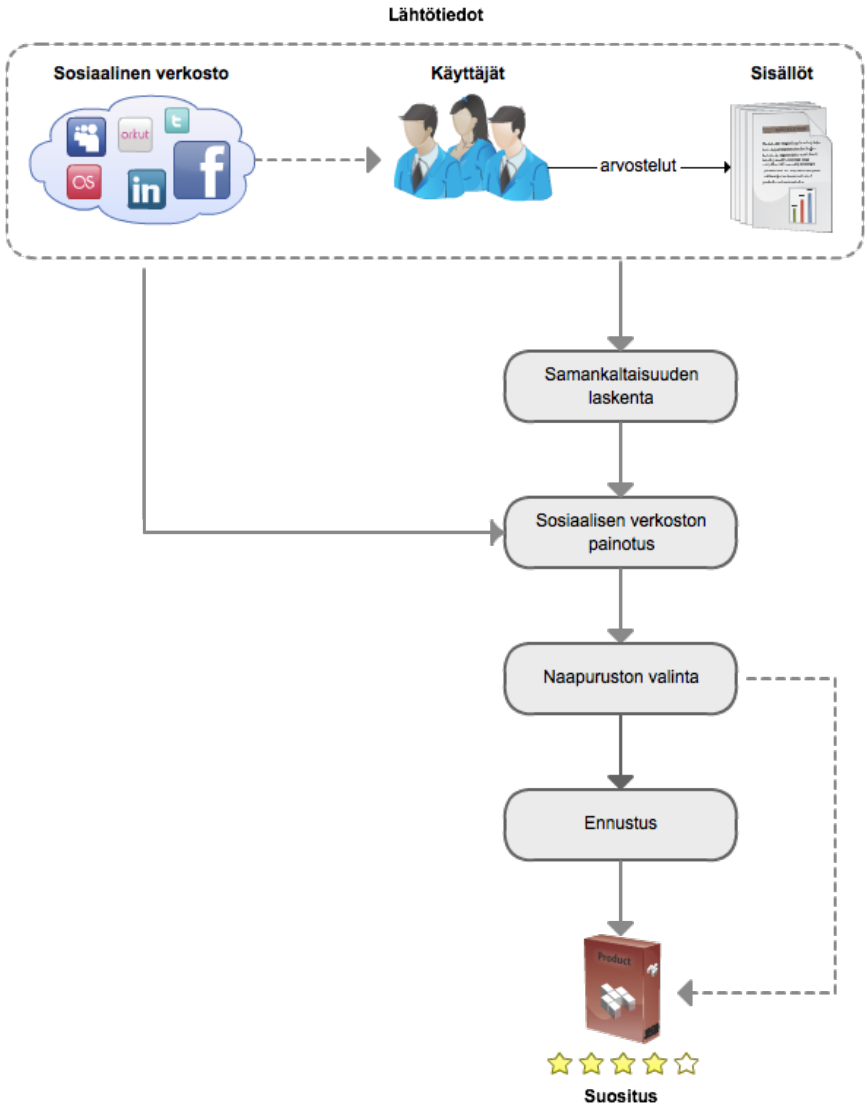
4 SOSIAALISEN VERKOSTON HYÖDYNTÄMINEN YHTEISÖLLISESSÄ SUODATUKSESSA

Tutkielman päätarkoitus on selvittää, voidaanko sosiaalista verkostoa hyödyntää suosittelutulosten parantamiseen. Päähypoteesi on se, että tutuilta henkilöiltä saadut suositukset ovat parempia kuin tuntemattomilta käyttäjiltä tulevat. Aikaisemmissa luvuissa on esitetty suosittelujärjestelmien ja sosiaalisten verkostojen teoriaa sekä sosiaalista suosittelua. Tässä luvussa esitetään tutkimus, jonka tulosten avulla pyritään vastaamaan tutkielman pääkysymykseen ja sen tarkentaviin kysymyksiin.

Seuraavissa luvuissa esitetään tutkimuksen asetelma, toteutus ja saadut tulokset.

4.1 Tutkimusasetelma

Tutkimuksen lähtökohtana on sosiaalisen verkoston yhdistäminen yhteisöllistä suodatusta käyttävään suosittelujärjestelmään. Yhdistetyn järjestelmän avulla tutkitaan, miten suosittelutulosten käy, jos niiden laskennassa annetaan enemmän painoarvoa sosiaalisilla suhteille kuin lasketuille korrelaatioarvoille. Kuvassa 4.1 on kuvattu tutkimuksen toteutusasetelma.



Kuva 4.1. *Tutkimuksen toteutusasetelma.*

Kuva 4.1 on mukailtu kuvan 2.3 yhteisöllisen suodatuksen prosessimallista. Samankaltaisuuden ja ennusteen laskennan voidaan ajatella olevan vain erilaisia komponentteja, joita voidaan vaihtaa suositustulosten parantamiseksi. Kuvassa 4.1. lisäkomponenttina on sosiaalisten ominaisuuksien painotus, jolla voidaan suosia sosiaalisen verkoston käyttäjiä naapuruston valinnassa. Kuten kuvassa 2.3, voidaan myös kuvan 4.1 mallissa naapuruston valinnan jälkeen joko laskea ennusteet, tai vain suositella samankaltaisia tuotteita tai käyttäjiä.

Suosittelujärjestelmän toteuttaminen on kuitenkin hyvin sovelluskohtaista ja laskennallisesti raskasta. Lisäksi tarvittavan tuote- ja käyttäjäkannan kerääminen vie aikaa miltä tahansa palvelulta. Koska valmista suosittelujärjestelmää ei ollut tarjolla, tutkimuksessa päätettiin hyödyntää jo olemassa olevaa tietoa. Suosittelujärjestelmiä kehittävät useat tutkimusryhmät, ja heidän keräämäänsä aineistoa on vapaasti saatavilla (esimerkiksi GroupLens Research 2012 tai Netflix 2012).

Tämän vuoksi tutkimuksen toteuttamiseksi päätettiin edetä analysoimalla jo olemassa olevan, yhteisöllistä suodatusta käyttävän suosittelujärjestelmän aineistoa. Tällä tavoin saatuun tietoon voidaan yhdistää sosiaalisesta verkostoitumispalvelusta saatua dataa, minkä avulla voidaan testata suositusten toimivuutta yhdistetyllä käyttäjäkannalla. Tutkimuksen analogia on, että olemassa oleviin palveluihin voidaan sosiaalisin liitännäisin yhdistää sosiaalisia ominaisuuksia ja tietoa sosiaalisista verkostoista. Tässä tapauksessa voidaan kuvitella, että tutkimuksen kohteena on elokuvasivusto, jossa on olemassa sisäinen käyttäjätietokanta, mutta ei sosiaalista verkostoa. Yhdistämällä käyttäjien sisäinen tunniste sosiaalisen verkoston (esimerkiksi Facebookin) tunnisteeseen, voidaan käyttäjien välille muodostaa sosiaalinen graafi näiden ulkoisten palvelujen kautta.

Sosiaalisen verkoston tuomiseen käytettiin Facebook-palvelua, koska sen käyttäjämäärä on suuri ja se tarjoaa kattavat rajapinnat sosiaalisen graafin ohjelmalliseen tuomiseen ja käsittelyyn. Toisin kuin monet muut palvelut, mallintaa sen ystäväverkosto käyttäjän omaa tosielämän sosiaalista verkostoa. Kehittäjän näkökulmasta Facebookin sosiaalisen verkoston saa vähällä vaivalla käyttöönsä ohjelmallisesti Graph API -rajapinnan kautta. Suositeltavien tuotteiden arvioiden ohjelmallista keräystä varten olisi kuitenkin tarvittu oma Facebook-sovelluksensa. Tällaisen sovelluksen tekemistä ei nähty tarpeelliseksi. Arviot kerättiin yksinkertaisesti kyselylomakkeella, joka lähetettiin kirjoittajan omalle Facebook-ystäväpiirille.

Kyselyssä Facebook-käyttäjät arvioivat tuotteita samalla asteikolla kuin olemassa olevassa aineistossakin on tehty. Yhdistämällä käyttäjä- ja arviotiedot voidaan simuloida tilannetta, jossa yhteisöllistä suodatusta käyttävän suosittelujärjestelmän käyttäjät tuntevat osin toisensa Facebook-verkoston kautta. Voidaan siis arvioida, saadaanko tällä tavoin aikaan parempia suosittelutuloksia. Tarkoitus ei ole kuitenkaan tehdä mahdollisimman tehokasta suosittelualgoritmia vaan arvioida menetelmän toteutettavuutta ja saada suuntaa-antavia tuloksia. Kyselytutkimuksen otanta ei lopulta ollut kovin suuri, mutta pääpaino onkin toteutuksen prototypoinnissa, ei kyselytutkimuksen tarkkuudessa. Näihin tutkimuskysymyksen kannalta olennaisiin osiin otetaan tarkemmin kantaa luvussa 5.

4.2 Toteutustekniikat

Toteutuksessa käytettiin hyväksi PHP-ohjelmointikieltä tiedonlouhintaan ja laskentaan sekä relaatiotietokantaa tiedon tallentamiseen. Seuraavissa aliluvuissa on esitetty tutkimuksessa käytetyt toteutustekniikat.

4.2.1 Relaatiotietokanta ja MySQL

Tietokannat ovat välttämättömiä vuorovaikutteisille www-sovelluksille, sillä niiden avulla voidaan erottaa toisistaan sivuston sisältö ja sen ulkoasu. Tietokannasta voidaan helposti hakea ja järjestellä halutut tiedot. Tietokantoja on erityyppisiä, kuten esimerkiksi hierarkkisia tietokantoja sekä relaatio- ja verkkotietokantoja.

Relaatiotietokanta on näistä yksinkertaisin ja joustavin. Se myös toteuttaa parhaiten tietokannalle asetettavat vaatimukset. Useimmat markkinoilla olevat tietokannat ovatkin juuri relaatiotietokantoja. Relaatiotietokanta koostuu useista relaatioista eli tauluista, joihin tieto tallennetaan riveinä ja sarakkeina. Tallennetut tiedot jaetaan relaatioihin siten, että yksi tieto tallennetaan vain yhteen paikkaan. Relaatiotietokantaan tallennetaan myös tietoja eri taulujen välisistä yhteyksistä. (Mustonen-Ollila 2006)

MySQL on nopea ja tehokas relaatiotietokanta. Se on avoimeen lähdekoodiin perustuva tietokannan hallintajärjestelmä, joka sopii erityisen hyvin yhteen Apache-palvelinympäristön ja PHP-ohjelmointirajapinnan kanssa. MySQL-tietokantaan tehdään hakuja SQL-kyselykielen avulla. SQL-kieli on standardoitu ja laajimmin käytetty kieli relaatiotietokantojen yhteydessä. SQL ei ole pelkkä kyselykieli, vaikka nimi siihen viittaakin. Kyselyjen tekemisen lisäksi sillä voidaan esimerkiksi muuttaa tietokannan rakennetta tai muokata kannan tietoja.

Tutkimusta varten tietokantaan luotiin neljä relaatiota, joihin tallennettiin tutkimuksessa käytetyt tiedot: käyttäjät, käyttäjien samankaltaisuus sekä elokuvat ja arvostelut. Taulukossa 4.1 on esitetty käyttäjärelaation rakenne. Relaation pääavain, eli tunniste, jolla jokainen rivi yksilöidään, on alleviivattu. Kenttien tietotyypit ovat MySQL-tietokannassa käytettyjä tyyppejä. Tietotyypin jälkeen suluissa on ilmoitettu kunkin tietotyypin vaatima tilanvaraus.

Taulukko 4.1. *Käyttäjärelaatio.*

user

Kenttä	**Tietotyyppi**	**Selite**
uid	varchar (25)	Käyttäjän tunniste, sisältää maksimissaan 25 merkkiä
title	varchar (255)	Käyttäjän nimi, maks. 255 merkkiä
isfb	int (1)	Kertoo, kuuluuko käyttäjä Facebook-verkostoon (0 = ei, 1 = kyllä)

Taulukossa 4.2 on esitetty relaatio, johon käyttäjien samankaltaisuudet tallennetaan. Relaatiossa on kaksi pääavainta, joista ensimmäinen (*uid1*) kuvaa aktiivista käyttäjää ja toinen (*uid2*) käyttäjää, jonka suhteen laskenta on tehty. Termin λ merkitys selitetään luvussa 4.4.

Taulukko 4.2. *Käyttäjien samankaltaisuus -relaatio.*

user_similarity

Kenttä	**Tietotyyppi**	**Selite**
uid1	varchar (25)	Käyttäjän 1 tunniste
uid2	varchar (25)	Käyttäjän 2 tunniste
cosine	float	Kosinietäisyydellä laskettu korrelaatio
cosine_10	float	Kosinietäisyys, $\lambda = 0.1$
cosine_20	float	Kosinietäisyys, $\lambda = 0.2$
cosine_30	float	Kosinietäisyys, $\lambda = 0.3$
cosine_40	float	Kosinietäisyys, $\lambda = 0.4$
cosine_50	float	Kosinietäisyys, $\lambda = 0.5$
cosine_60	float	Kosinietäisyys, $\lambda = 0.6$
cosine_70	float	Kosinietäisyys, $\lambda = 0.7$
cosine_80	float	Kosinietäisyys, $\lambda = 0.8$
cosine_90	float	Kosinietäisyys, $\lambda = 0.9$
cosine_100	float	Kosinietäisyys, $\lambda = 1$

Taulukossa 4.3 on esitetty elokuvarelaatio.

Taulukko 4.3. *Elokuvarelaatio.*

movie

Kenttä	**Tietotyyppi**	**Selite**
id	int (11)	Elokuvan tunniste
title	varchar (255)	Elokuvan nimi
year	int (4)	Elokuvan ilmestymisvuosi
genres	varchar (255)	Elokuvan luokittelutermit
imdb	float	Elokuvan arvosana IMDB-palvelussa

Taulukossa 4.4 on esitetty arvostelurelaatio.

Taulukko 4.4. *Arvostelurelaatio.*

movie_rating

Kenttä	**Tietotyyppi**	**Selite**
uid	varchar (25)	Käyttäjän tunniste
movie	int (11)	Elokuvan tunniste
rating	int (1)	Elokuvan arvosana välillä 1…5

Relaatioihin tallennettavista tiedoista kerrotaan tarkemmin luvussa 4.4.

4.2.2 PHP

PHP on avoimeen lähdekoodiin perustuva palvelinpohjainen ohjelmointikieli. Palvelinpohjainen ohjelmisto tarkoittaa sitä, että koodi suoritetaan palvelimella, jolloin se ei vaadi mitään erityistä tukea selaimelta. Palvelinpohjaisuuden ansiosta PHP-sivujen avulla voidaan muun muassa käsitellä palvelimella olevia tiedostoja, ajaa ohjelmia komentotulkin avulla sekä käyttää erilaisia tietokantoja.

PHP-sivut suoritetaan palvelimelle asennettavan PHP-tulkin avulla. PHP ei ole varsinainen ohjelmointikieli, sillä sitä ei käännetä binäärimuotoon. PHP onkin niin sanottu skriptikieli, joka tulkataan PHP-tulkin avulla joka suorituskerralla uudestaan. PHP-koodia voidaan kirjoittaa HTML-merkkauksen yhteyteen. Se kuitenkin erotetaan muusta sisällöstä erityisellä merkinnällä. Toisin kuin tavallinen HTML-sivu, PHP-koodia ei lähetetä suoraan asiakassovellukselle, vaan se esikäsitellään PHP-ohjelman toimesta. Tiedostossa olevat HTML-elementit jätetään käsittelemättä, mutta PHP-koodi tulkataan ja suoritetaan. (Zandstra 2001)

4.2.3 JSON

JSON (*JavaScript Object Notation*) on kevyt, tekstipohjainen datan esitysmuoto verkkosovellusten välisen tiedon välittämiseen (JSON 2012). Se perustuu verkkosivuilla käytettyyn JavaScript-skriptikieleen. Tarkoitus on, että JSON-muotoon kirjoitettua dataa on sekä ihmisen helppo lukea että koneen tulkita. Käytännössä JSON-esitysmuoto vastaa JavaScript-kielen oliota ja se voidaan tulkata JavaScript-ohjelmassa suoraan olioksi. JSON on kuitenkin kieliriippumaton ja sen käsittelemiseen on saatavilla useita toteutuksia kaikille yleisimmille ohjelmointikielille. Kieliriippumattomuuden ansiosta JSON on hyvä tapa välittää erilaisia ohjelmallisia tietorakenteita ja dataa verkkosovellusten välillä.

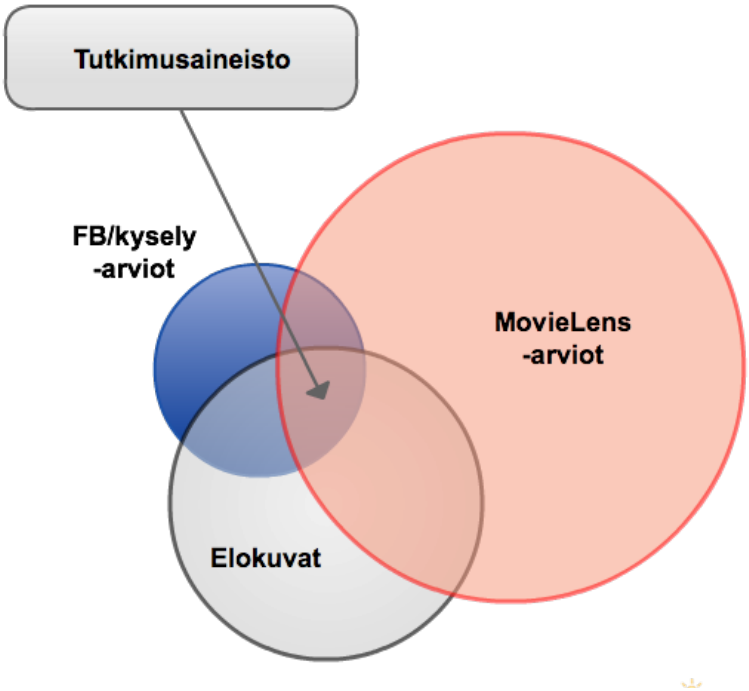
4.3 Aineisto

Tutkimuksen aineisto koostuu elokuvien arvostelutiedoista. Tutkimuksen kohteeksi valikoitui elokuvat kahdesta syystä. Ensinnäkin elokuvat ovat monelle hyvin henkilökohtainen asia, ja usein ihmiset myös määrittävät persoonaansa elokuvien kautta. Nykyajan popkulttuuri on värittynyt paljolti elokuvien kautta. Lähes kaikki nykyaikaisessa länsimaalaisessa yhteiskunnassa elävät katselevat elokuvia säännöllisesti. Elokuva on arvostelukohteena atominen, eli toisin kuin musiikkilevy, se ei koostu useasta, joskus eriarvoisestakin osasta. Lisäksi elokuva on tarpeeksi pitkä, jotta siitä ehtii muodostua katsojalle vankka mielipide. Elokuvia arvostellaan usein viiden tähden skaalalla, joka on tarpeeksi kuvaava yksikkö suosittelujärjestelmää varten.

Vähemmän teoreettisella tasolla toinen syy elokuvien valintaan oli saatavilla oleva valmis tutkimusaineisto ja olemassa olevat tutkimukset. Elokuvien suositteluun on viime aikoina keskitytty paljon esimerkiksi vuonna 2009 päättyneen Netflix-kilpailun myötä (Netflix 2012). Kilpailussa elokuvien vuokraukseen keskittyvä Netflix-

verkkopalvelu lupasi miljoona dollaria ensimmäiselle tutkimusryhmälle, joka parantaa heidän olemassa olevaa suosittelujärjestelmäänsä kymmenellä prosentilla aikaisempaan toteutukseen verrattuna. Eräs suosittu, myöskin elokuviin liittyvä tutkimuskohde on MovieLens-suosittelupalvelu. Palvelun on toteuttanut tietojärjestelmiin ja suosittelujärjestelmiin keskittynyt tutkimusryhmä ja se myös tarjoaa tutkijoiden käyttöön vapaasti palvelusta saatuja arvostelutietoja (GroupLens Research 2011).

Tutkielmassa käytettiin hyväksi MovieLens-palvelun dataa. Saatu data on ”nimetöntä”, eli palvelun käyttäjät eivät tunne toisiaan ja arvostelevat itsenäisesti elokuvia. Tähän aineistoon lisätään Facebookista, jossa käyttäjät siis tuntevat toisensa, saatua kyselytietoa. Kyselytutkimus tehtiin Facebookissa marraskuun 2011 aikana ja tutkimuskohteena oli kirjoittajan oma ystäväpiiri. Kuten kuvassa 4.2 on esitetty, koostuu tutkimusaineisto siis valikoiduista elokuvista sekä MovieLens- ja Facebook-käyttäjien arvioista näihin elokuviin.



Kuva 4.2. *Tutkimusaineisto.*

Seuraavissa luvuissa on esitetty tarkemmin aineiston sisältö.

4.3.1 MovieLens-palvelun aineisto

MovieLens on elokuvien suositteluun keskittyvä palvelu (GroupLens Research 2012). Palvelu pyrkii tarjoamaan käyttäjälle heti toimivia suosituksia. Uuden käyttäjän onkin rekisteröintilomakkeen yhteydessä arvosteltava 15 satunnaisesti valittua elokuvaa. Tällä pyritään estämään aikaisemmin mainitut uuden käyttäjän ongelmat. Palvelussa käytetään yhteisöllistä suodatusta, eli järjestelmä vertaa käyttäjän pisteytyksiä toisten käyttäjien pisteisiin ja muodostaa niiden perusteella ennusteet elokuville, joita käyttäjä

ei ole vielä pisteyttänyt. Suositusten tarkkuus lisääntyy sitä mukaa kuin käyttäjä arvioi näkemiään elokuvia.

MovieLens-palvelun ja suositusalgoritmin on kehittänyt GroupLens-tutkimusryhmä. Tutkimusryhmän palvelusta koostamat datasarjat ovat vapaasti saatavilla tutkimusryhmän verkkosivuilta (GroupLens Research 2011). Palvelusta kerätty data on tieteellisen yhteisön vapaasti käytettävissä. Saatavilla on kolme erikokoista tietosarjaa. Kohteeksi valittiin ”MovieLens 1M” -sarja, joka käsittää miljoona arvostelua 4000:sta elokuvasta 6000:lta käyttäjältä. Tietosarja koostuu kolmesta tiedostosta, joissa ovat erikseen määriteltynä elokuvat, käyttäjät ja arviot. Jokaisessa tiedostossa yksi rivi vastaa yhtä tietuetta. Taulukossa 4.5 on esitetty tiedostot, niiden sisältämä tieto ja tietueiden määrät.

Taulukko 4.5. *MovieLens 1M -tietosarja.*

Tiedosto	**Selite**	**Lukumäärä**
movies.dat	Elokuvat	3884
ratings.dat	Arviot	1000210
users.dat	Käyttäjät	6040

Näistä tiedostoista tutkimuksen kannalta olennaisia ovat elokuvat ja arviot sisältävät tiedostot. Arviotietokanta sisältää jokaiselle käyttäjälle yksilöivän tunnisteen. Tarkempia tietoja käyttäjästä ei tässä yhteydessä tarvita. Käyttäjätiedosto sisältää kuitenkin tarvittaessa demografista tietoa kustakin käyttäjästä, kuten sukupuoli ja ikä. Elokuvatiedoston tietueet on koodattu siten, että kukin tiedoston rivi vastaa yhtä elokuvatietuetta ja kullakin rivillä elokuvien tiedot on erotettu toisistaan kahdella kaksoispisteellä. Elokuvatietueen formaatti on seuraavanlainen:

```
MovieID::Title::Genres
```

jossa *MovieID* on elokuvan yksilöivä tunniste, *Title* kertoo elokuvan nimen ja *Genres* on lista elokuvan aihetyypeistä. Aihetyyppien listan arvot on erotettu toisistaan putkimerkillä ”|”, mutta elokuvien aihetyypeillä ei ole merkitystä tutkimuksen kannalta, joten saatu tekstirivi tallennettiin sellaisenaan tietokantaan. Listauksessa 4.1 on esimerkki elokuvatiedoston sisällöstä.

Listaus 4.1. *Esimerkki elokuvatiedoston sisällöstä.*

```
1::Toy Story (1995)::Animation|Children's|Comedy
2::Jumanji (1995)::Adventure|Children's|Fantasy
3::Grumpier Old Men (1995)::Comedy|Romance
4::Waiting to Exhale (1995)::Comedy|Drama
5::Father of the Bride Part II (1995)::Comedy
```

Arviotiedoston sisältämä tieto on koodattu samoin kuin elokuvatiedostonkin, mutta tietueet on merkitty seuraavasti:

```
UserID::MovieID::Rating::Timestamp
```

jossa *UserID* on käyttäjän yksilöivä tunniste, *MovieID* elokuvan tunniste, *Rating* on käyttäjän antama arvio elokuvalle ja *Timestamp* aikaleima, jolloin arvio on annettu. Käyttäjien antamat arviot ovat välillä 1-5, jossa 1 tarkoittaa huonoa ja 5 hyvää elokuvaa. Arviot tallennettiin aikaleimaa lukuun ottamatta tietokannan relaatioon. Listauksessa 4.2 on esimerkki arviotiedoston sisällöstä.

Listaus 4.2. *Esimerkki arviotiedoston sisällöstä.*

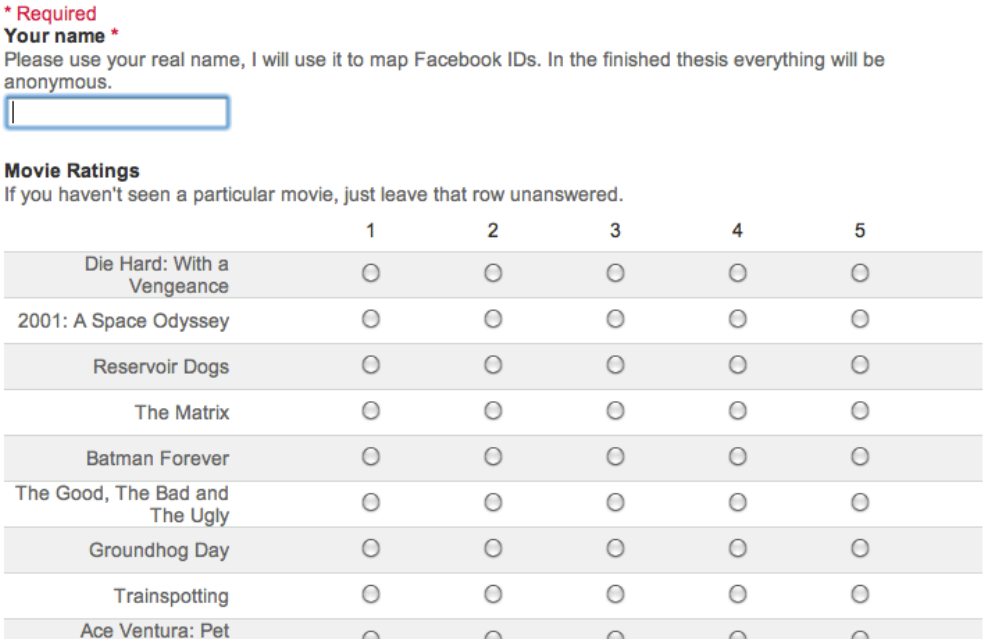
```
1::661::3::978302109
1::914::3::978301968
1::3408::4::978300275
1::2355::5::978824291
```

Käyttäjätiedostoa ei tarvittu, joten sen sisältö jätettiin käsittelemättä.

4.3.2 Kyselytutkimuksen tulokset

Kyselytutkimus toteutettiin marraskuun 2011 aikana. Kyselyä varten MovieLens-aineiston elokuvien joukosta valittiin arvosteluun satunnaisesti 50 elokuvaa. Elokuvien valinnan kriiterinä oli se, että ne löytyivät myös MovieLens-palvelun valikoimasta sekä se, että ne olivat kohtuullisen tunnettuja. Joukossa oli sekaisin elokuvia eri lajityypeistä sekä laadultaan eritasoisia filmejä. Tarkka lista valituista elokuvista on liitteessä 1.

Kyselylomake toteutettiin Google Docs -palvelun avulla, joka mahdollisti kyselylomakkeen helpon toteutuksen. Tämä toteutus valittiin siksi, että käyttäjien vastaukset tallentuivat automaattisesti Google Docs -palveluun omana taulukkotiedostona. Tästä taulukkotiedostosta oli helppo ottaa talteen ja erottaa kyselyn tulokset. Kyselylomakkeessa kysyttiin käyttäjän nimeä sekä arvioita 50:lle valitulle elokuvalle. Arvosteluskaala oli sama kuin MovieLens-palvelussa eli 1-5, jossa 1 kuvaa huonoa elokuvaa ja 5 hyvää. Jos vastaaja ei ollut nähnyt jotakin elokuvaa, häntä pyydettiin jättämään kyseinen rivi tyhjäksi. Lomake kirjoitettiin englannin kielellä, jotta kaikki kirjoittajan ulkomaalaisetkin Facebook-kontaktit voisivat vastata kyselyyn. Kuvassa 4.3 on ote kyselylomakkeesta.



* Required
Your name *
Please use your real name, I will use it to map Facebook IDs. In the finished thesis everything will be anonymous.

Movie Ratings
If you haven't seen a particular movie, just leave that row unanswered.

	1	2	3	4	5
Die Hard: With a Vengeance	○	○	○	○	○
2001: A Space Odyssey	○	○	○	○	○
Reservoir Dogs	○	○	○	○	○
The Matrix	○	○	○	○	○
Batman Forever	○	○	○	○	○
The Good, The Bad and The Ugly	○	○	○	○	○
Groundhog Day	○	○	○	○	○
Trainspotting	○	○	○	○	○
Ace Ventura: Pet	○	○	○	○	○

Kuva 4.3. *Ote kyselytutkimuksen lomakkeesta.*

Kyselytutkimuslinkki jaettiin Facebookissa kirjoittajan omalle ystäväpiirille, jota pyydettiin vastaamaan kyselyyn. Linkkiä jaettiin sekä tilapäivityksenä kirjoittajan omalla seinällä että yksityisviestinä. Kysely tavoitti lopulta 21 vastaajaa 161:stä silloisesta kontaktista. Kyselyn kattavuus oli siis noin 13% kirjoittajan kaikista kontakteista. Vastausten määrä katsottiin riittäväksi, koska tarkoitus ei ollut tehdä täysin kattavaa tutkimusta. Pääpaino tutkimuksella onkin tarkkuuden sijaan sosiaalisen verkostotiedon yhdistämisessä suosittelujärjestelmän toimintaan. Tarkoitus on prototypoida menetelmiä ja arvioida niitä sekä numeerisesti että laadullisesti. Vertailu on joka tapauksessa hyvin subjektiivista, sillä näkökulma on vain kirjoittajan omiin kokemuksiin ja ystäväpiiriin kohdistuva.

4.3.3 Facebook-käyttäjien tiedot

Facebook-verkosto haettiin Graph API -rajapintakutsulla. Graph API -rajapinta toimii HTTP-protokollan välityksellä, joten sitä voidaan käyttää myös tavallisella verkkoselaimella. Esimerkiksi menemällä osoitteeseen https://graph.facebook.com/*tunniste* (*tunniste* on käyttäjän yksilöivä tunniste), saa esille käyttäjän omat tiedot. Rajapinta palauttaa HTTP-pyynnön vastauksena JSON-muotoisen tietueen, jossa käyttäjän tiedot on määritelty. Jokaisella sosiaalisen graafin toimijalla (esimerkiksi ryhmät, tapahtumat tai tilapäivitykset) on oma tunnisteensa, ja näiden tietojen haku onnistuu samalla tavoin. Listauksessa 4.3 on esimerkki JSON-tietueesta, jossa on esitetty kuvitellun henkilön Facebook-sivun tiedot. Jos kyse olisi oikeasta henkilöstä, olisivat tiedot saatavilla osoitteesta https://graph.facebook.com/*erkki.esimerkki*, jossa *erkki.esimerkki* on siis kuvitellun henkilön tunniste.

Listaus 4.3. *Kuvitellun henkilön Facebook-sivun tiedot JSON-muodossa.*

```
{
   "id": "3453678642",
   "name": "Erkki Esimerkki",
   "first_name": "Erkki",
   "last_name": "Esimerkki",
   "username": "erkki.esimerkki",
   "locale": "fi_FI"
}
```

Saatavilla ovat ne tiedot, joihin rajapintaa käyttävällä kehittäjällä olisi itse palvelussakin pääsy. Tämä tarkoittaa sitä, että verkoston ulkopuolisten käyttäjien tiedoista nähtävissä on vain julkisiksi määrätyt asiat. Facebookin sosiaalisen graafin toimijoiden suhteiden tarkastelu on mahdollista osoitteessa https://graph.facebook.com/*tunniste*/*yhteys*, jossa *tunniste* on käyttäjän yksilöivä tunniste ja *yhteys* on halutun solmun tyyppi, esimerkiksi käyttäjän kontaktit. Tämän lisäksi pyynnön yhteydessä on välitettävä sovellusavain (*access token*), jolla palvelu rajaa pääsyä vain käyttäjän oman sosiaalisen graafin tietoihin.

Tutkimusta varten tarvittiin lista siihen osallistuneista käyttäjistä. Graph API -rajapinnasta noudettiin lista kaikista kirjoittajan Facebook-ystävistä. Listauksessa 4.4 on esitetty esimerkki palautetun listan sisällöstä. Käyttäjien oikeita nimiä tai tunnisteita ei voida tässä esittää yksityisyyden suojaamiseksi, joten esimerkin arvot ovat keksittyjä.

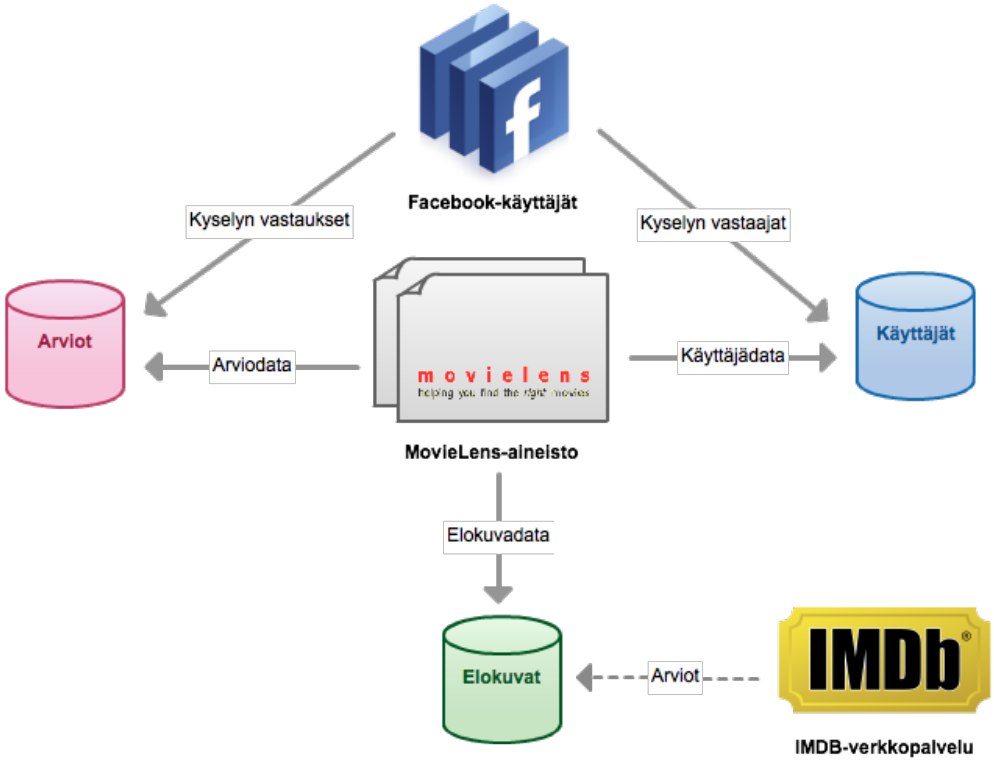
Listaus 4.4. *Facebook-ystävien lista JSON-muodossa.*

```
{
    "data": [
        {
            "name": "Etunimi1 Sukunimi1",
            "id": "1111111"
        },
        {
            "name": "Etunimi2 Sukunimi2",
            "id": "222222"
        },
        // ...
        {
            "name": "Etunimi3 Sukunimi3",
            "id": "3333333"
        }
    ]
}
```

Saatu lista koostui 161 tietueesta. Joukosta karsittiin pois kaikki muut paitsi kyselytutkimukseen osallistuneet henkilöt. Tämän jälkeen käyttäjien tiedot tallennettiin tietokannan käyttäjärelaatioon siten, että käyttäjät voitiin jälkikäteen erottaa joko Facebook- tai MovieLens-käyttäjiin.

4.3.4 Koottu aineisto

Kuvassa 4.4 on esitetty, miten MovieLens-palvelun ja kyselytutkimuksen aineisto on yhdistetty.



Kuva 4.4. *Eri palveluista koottu tutkimusaineisto.*

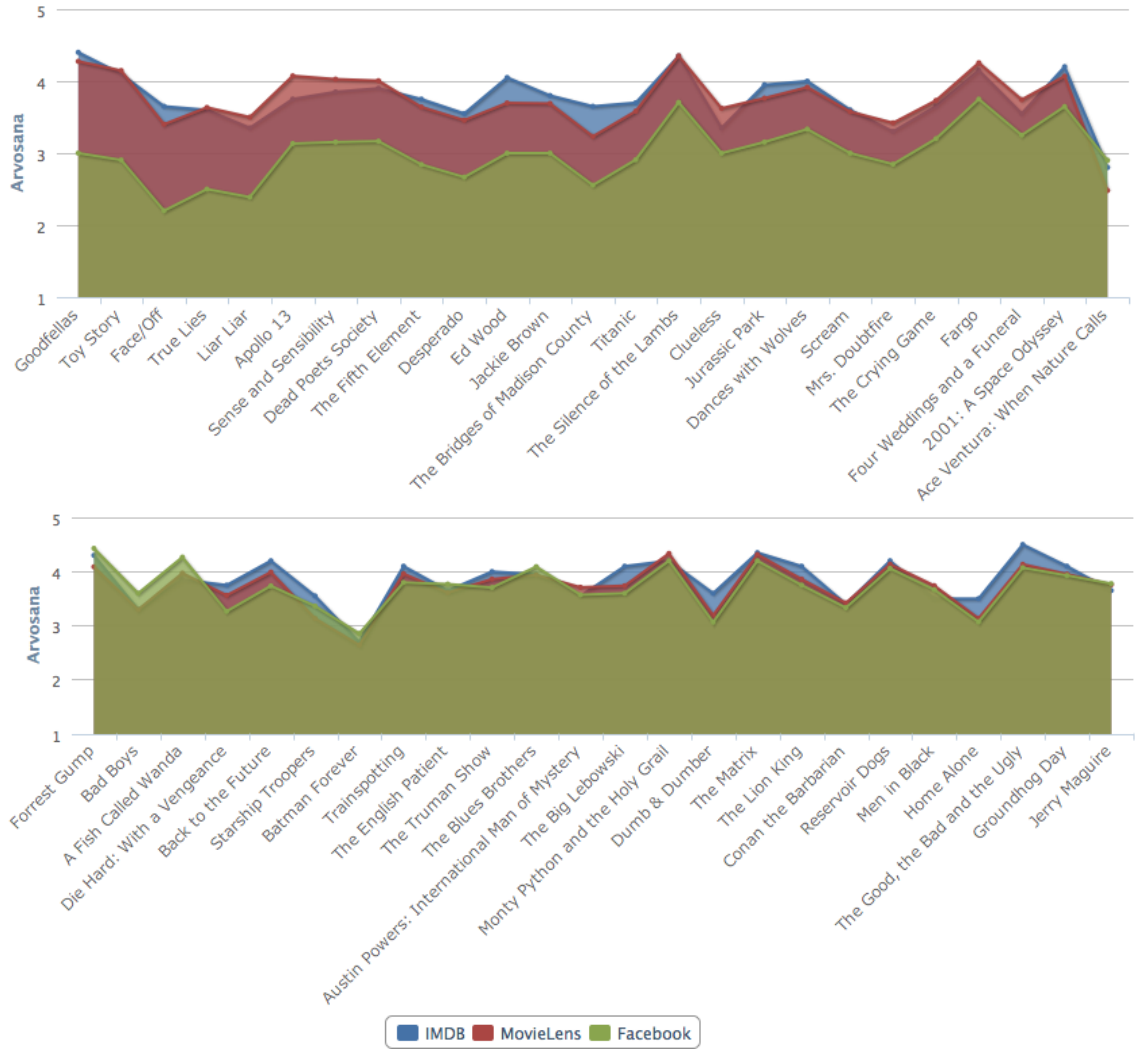
MovieLens-elokuvatiedostosta luettiin vain ne elokuvat, jotka valittiin mukaan kyselyyn. Mukaan valittiin alun perin 50 elokuvaa, mutta yksi niistä jouduttiin poistamaan, koska sitä ei ollut määritetty MovieLens-tiedostossa lainkaan. Tämän virheen vuoksi käsiteltyjä elokuvia oli lopulta vain 49. Elokuvien tiedot tallennettiin omaan relaatiotietokannan tauluunsa. Elokuvien tietoihin lisättiin myös vertailun vuoksi IMDB-elokuvasivuston arvio, joka perustuu sivuston käyttäjäyhteisön mielipiteeseen. Arviotiedostosta luettiin ne rivit, jotka vastasivat valittuja elokuvia. Arviot tallennettiin omaan tauluunsa relaatiotietokannassa. Tämän jälkeen elokuvien arvioiden joukosta yksilöitiin arviot antaneet käyttäjät, jotka tallennettiin omaan relaatioonsa.

Tietokannan arvio- ja käyttäjärelaatioon lisättiin tämän jälkeen kyselystä saadut arviot sekä kyselyyn vastanneiden käyttäjien tunnisteet. MovieLens- ja Facebook-käyttäjät tallennettiin samaan relaatioon, mutta erotettiin toisistaan erillisellä tunnistekentällä. Tämän avulla voidaan tunnistaa kumpaan joukkoon arvostelun antanut käyttäjä kuuluu. Taulukossa 4.6 on esitetty tutkimuksessa käytetyn lopullisen aineiston koko. Käyttäjien ja arvioiden määrän kohdalla ”Facebook”-merkintä tarkoittaa sekä kyselyyn osallistuneita käyttäjiä että kirjoittajaa itseään. ”MovieLens”-merkintä tarkoittaa niiden käyttäjien joukkoa, jotka ovat arvostelleet valitut elokuvat.

Taulukko 4.6. *Tutkimusaineiston koko.*

Aineisto	**Lukumäärä**
Elokuvat	49
Käyttäjät	5687
MovieLens	5665
Facebook	22
Arviot	64384
MovieLens	63615
Facebook	769

Kuvassa 4.5 on esitetty kyselyn tuloksina saadut arvosanojen keskiarvot. Samassa kuvassa on myös vertailun vuoksi arvosanojen MovieLens-datasta lasketut keskiarvot sekä IMDB-palvelun arviot elokuvista. Arvosanat on järjestetty niin, että ensin ovat elokuvat, joiden MovieLens- ja Facebook-arviot eroavat eniten toisistaan.



Kuva 4.5. *Arvosanajakaumat MovieLens- ja IMDB-palveluista sekä kyselytutkimuksesta.*

IMDB:n arvosteluskaala on normaalisti välillä 0-10, mutta kuvassa 4.5 IMDB-arviot on skaalattu samalle arvosteluvälille kuin muukin data. Kuvasta selviää, että MovieLens- ja IMDB-palveluiden arvosanat vastaavat paljolti toisiaan, kun taas kyselystä saadut arviot poikkeavat hieman tästä linjasta. Luultavasti tämä johtuu otannan pienuudesta, mutta poikkeamat eivät ole suuria.

4.4 Sosiaalisen verkoston painotus

Olennainen osa tutkimusta on se, miten tietoa sosiaalisista suhteista voidaan hyödyntää automaattisessa suosittelussa. Tutkimuksen lähteenä on käytetty Yuanin et al. (2009) tekemää tutkimusta, jossa he hyödynsivät sekä tietoa käyttäjien välisistä ystävyyssuhteista (*friendship*) että heidän suhteistaan eri yhteisöihin (*membership*) musiikin suosittelussa. Tämän tutkimuksen kannalta tieto jäsenyydestä ei ole olennainen, sillä tarkoitus on tutkia vain eksplisiittisen sosiaalisen suhteen vaikutusta suosittelutulokseen.

Ystävyyssuhdetta voidaan kuvata luvussa 3.2 esitetyn kaksiarvoisen sosiomatriisin avulla. Sosiomatriisia käyttäen voidaan muodostaa funktio, joka kertoo ovatko kaksi käyttäjää toistensa kontakteja. Tutkimusaineiston käyttäjien välinen suhde saadaan selville määritelmässä 4.1 esitetyn funktion avulla. Yhteydet käyttäjien välillä ovat kaksisuuntaisia.

Määritelmä 4.1. *Käyttäjien välinen suhde.*

$$f_{FB}(u,v) = \begin{cases} 1, \text{ jos } u \text{ ja } v \text{ ovat toistensa FB-kontakteja ja } u \neq v \\ 0, \text{ muulloin} \end{cases}$$

Kuten kuvassa 4.1 on esitetty, tapahtuu sosiaalisten suhteen painotus hyödyntämällä aikaisemmin laskettuja samankaltaisuusarvoja. Samankaltaisuus voidaan laskea samoilla menetelmillä kuin luvussa 2 on esitetty, tai käyttämällä jotain muuta menetelmää. Yuan et al. (2009) käyttivät omassa tutkimuksessaan määritelmän 2.1 mukaista kosinietäisyyttä, jonka arvot laskettiin sosiomatriisin käyttäjävektoreista. Määritelmässä 4.2 on esitetty kaava sosiaalisen verkoston painotusta varten, joka on mukailtu Yuanin et al. (2009) tutkimuksesta.

Määritelmä 4.2. *Sosiaalisen verkoston yhdistäminen samankaltaisuuden laskentaan. Kaavassa muuttuja λ on painoarvo, jonka samankaltaisuusarvo saa laskennassa ja $s_{u,v}$ on valittu samankaltaisuusarvojen joukko.*

$$\hat{s}_{u,v} = \lambda * s_{u,v} + (1 - \lambda) * f_{FB}(u,v)$$

Määritelmän 4.2 antama samankaltaisuusarvo painottaa sosiaalista verkostoa sitä enemmän, mitä suurempi λ on. Muuttujan λ arvo annetaan välillä 0…1, jossa arvolla 0 painotetaan täysin sosiaalista suhdetta ja arvolla 1 käytetään ainoastaan laskettua samankaltaisuutta. Tässä tutkimuksessa sosiaalisen samankaltaisuuden laskennassa on

käytetty funktiota $f_{FB}(u,v)$, koska laskenta perustuu vain yhden käyttäjän näkökulmaan eikä muita sosiaalisia suhteita ole järjestelmässä. Tulos olisi joka tapauksessa sama kuin kosinietäisyyttä käyttämällä.

Samankaltaisuuden laskentaan voidaan tehdä myös muita painotuksia. Esimerkiksi Yuanin et al. (2009) tutkimuksessa kolmantena painoarvona on käyttäjien jäsenyys eri musiikkiyhteisöissä ja niistä lasketut samankaltaisuudet. Golbeck (2006) on tutkinut käyttäjien eksplisiittisesti määrittämää luottamusta sosiaalisen verkostotiedon lisänä suosittelussa, kun taas Huang et al. (2010) ovat tutkineet käyttäjien välistä vaikutusvaltaa osana suosittelua.

Tässä tutkielmassa pyritään selvittämään, onko suosittelutuloksia mahdollista parantaa ottamalla käyttöön tieto yksinkertaisista sosiaalista suhteista. Tarkoitus ei ole tehdä mahdollisimman tehokasta suosittelualgoritmia, vaan tutkia toteutuksen helppoutta ja toteutettavuutta sekä numeerisesti että subjektiivisesti. Tämän vuoksi laskennassa käytettiin kosinietäisyyttä samankaltaisuuden laskentaan, koska se on yksinkertaisin laskutapa korrelaatiolle.

4.5 Tutkimuksen tulokset

Tässä luvussa esitetään tutkimuksen tulokset.

4.5.1 Käyttäjien samankaltaisuus

Vaikka sosiaalista verkostoa käytetään suosittelun tukena, ei voida suoraan sanoa, että verkoston käyttäjät ovat samankaltaisia. Kyseessä on pikemminkin oletus siitä, että käyttäjä luottaa enemmän tuntemaansa ihmiseen kuin tuntemattomaan suosittelujen yhteydessä. Tästä syystä on asiallista tarkistaa, kuinka suuri osuus sosiaalisten verkostojen jäsenillä on samankaltaisten käyttäjien joukossa.

Käyttäjien samankaltaisuutta verrattiin laskemalla käyttäjille samankaltaisuusarvo määritelmän 2.1 mukaisesti. Listauksessa 4.5 on esitetty laskennassa käytetty PHP-funktio. Funktiossa kerätään aktiivisen ja verrattavan käyttäjän arvostelut taulukkoihin, jotka kerrotaan keskenään määritelmän 2.1 mukaisesti. Funktio *dot()* laskee kahden vektorin (tässä tapauksessa taulukon) pistesumman ja funktio *magnitude()* vektorien etäisyyden. PHP-kielestä ei löytynyt tukea näille, joten funktiot toteutettiin itse.

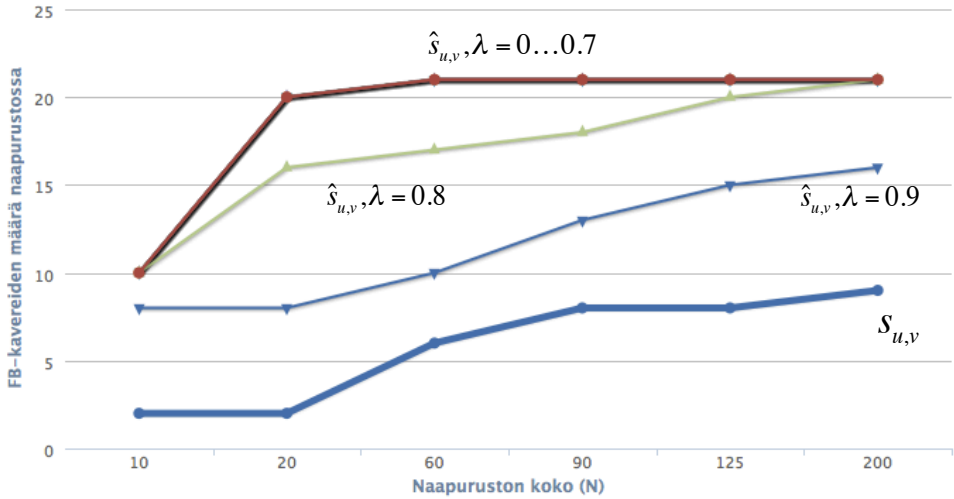
Kuten listauksesta 4.5 selviää, tehdään samankaltaisuuden laskennan yhteydessä monta tietokantakyselyä. Todellisessa järjestelmässä raskaita tietokantakyselyitä pyrittäisiin välttämään, mutta yksinkertaisuuden vuoksi toteutus tehtiin näin.

Listaus 4.5. *Kosinietäisyyden laskeva PHP-funktio.*

```
function sim_cosine($u1, $u2, $movies) {
    $u1_ratings = array();
    $u2_ratings = array();

    foreach($movies as $movie) {
        $u1_ratings[] = (int)db_get_value("SELECT rating FROM
movie_rating WHERE uid = '$u1' AND movie = $movie");
      $u2_ratings[] = (int)db_get_value("SELECT rating FROM
movie_rating WHERE uid = '$u2' AND movie = $movie");
    }

    return (dot($u1_ratings, $u2_ratings) /
(magnitude($u1_ratings) * magnitude($u2_ratings)));
}
```

Painotettu samankaltaisuus laskettiin määritelmän 4.2 mukaisesti muuttujan λ eri arvoilla. Samankaltaisuus laskettiin painottamalla sosiaalista verkostoa 10% välein, eli laskennassa muuttuja λ sai siis arvoja 0.1:n välein. Naapuruston kokoa vertailtiin erilaisilla otannoilla aina kymmenestä käyttäjästä kahteen sataan. Vertailua varten naapurustojen koot valittiin samoiksi kuin Yuanin et al. (2009) tutkimuksessa. Pienet naapurustot ovat enemmän käytössä todellisissa järjestelmissä, sillä laskenta usean naapurin suhteen hidastaa laskentaa (Su & Khoshgoftaar 2009).

Elokuvat jaettiin satunnaisotannalla puoliksi kahteen osaan – opetusaineistoon ja testiaineistoon. Opetusaineistoon kuuluu siis 25 elokuvaa, joiden avulla laskenta suoritetaan. Loppuosaa käytetään laskennan testaamiseen. Laskennassa naapurustojen koot ovat 10, 20, 60, 90, 125 ja 200. Samankaltaisten käyttäjien lukumäärät eri naapurustoilla on esitetty kuvassa 4.6.



Kuva 4.6. *Facebook-kontaktien määrät samankaltaisten käyttäjien joukossa eri naapurustoilla.*

Kuvasta 4.6 on nähtävissä, että puhtaalla kosinietäisyydellä laskettuna samankaltaisten Facebook-kontaktien määrä ei ylitä kymmentä edes 200 käyttäjän naapurustolla. Painottamalla sosiaalista verkostoa Facebook-kontaktien määrä luonnollisesti nousee. Tästä on nähtävissä se, että sosiaalisen verkoston jäsenten arviot eivät korreloi keskenään, mikä on tietysti normaalia, sillä ihmisillä on erilaisia mielipiteitä.

4.5.2 Suosittelutulokset

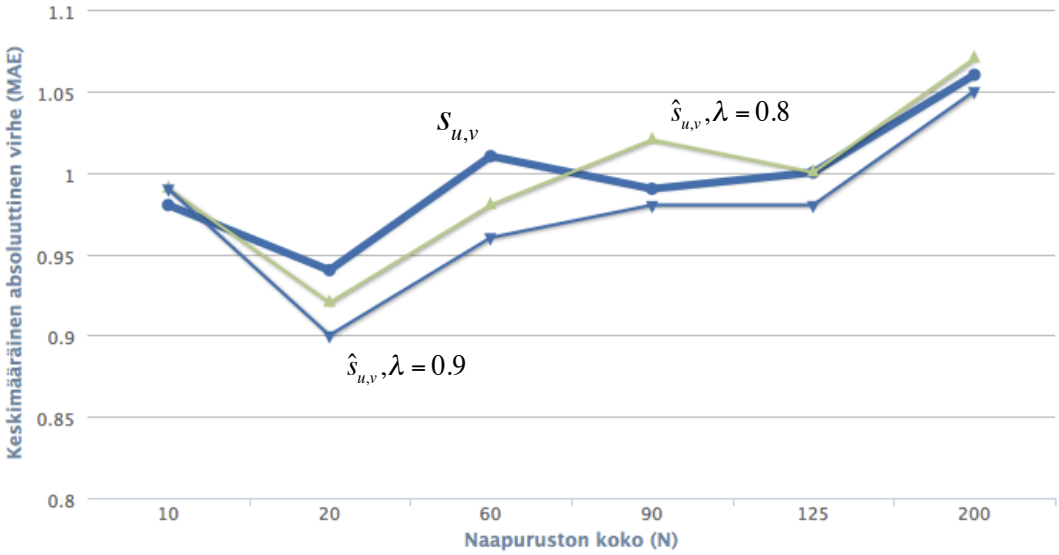
Tässä luvussa esitetään tulokset siitä, miten suosittelutulokset muuttuvat sitä mukaa kuin sosiaalista verkostoa painotetaan laskennassa. Suosittelujärjestelmän tarkkuutta voidaan mitata vertaamalla ennustetta annettuun oikeaan arvoon. Laskettuja samankaltaisuuksia hyväksi käyttäen laskettiin ennusteet elokuvasarjan testiosalle. Samankaltaisuudet laskettiin ensin käyttämällä sekä pelkkää kosinietäisyyttä että painottamalla sosiaalisesta verkostoa 10% välein. Arvot tallennettiin käyttäjien samankaltaisuus –relaatioon. Tämä tehtiin siis elokuvalistan toiselle puolikkaalle eli testiaineistolle.

Ennusteet laskettiin määritelmän 2.5 mukaisesti kaikilla samankaltaisuuksien painotuksilla ja erikokoisilla naapurustoilla. Listauksessa 4.6 on esitetty PHP-funktio, jota käyttäen ennuste laskettiin. Funktion parametri *$sim_f* kertoo, mitä samankaltaisuusarvoa kulloinkin käytetään. Kyseessä on siis joko puhdas kosinietäisyys tai sosiaalisella suhteella painotettu arvo ja se saadaan käyttäjien samankaltaisuus -relaatiosta, johon se edellä tallennettiin.

Listaus 4.6. *Ennusteen laskeva PHP-funktio.*

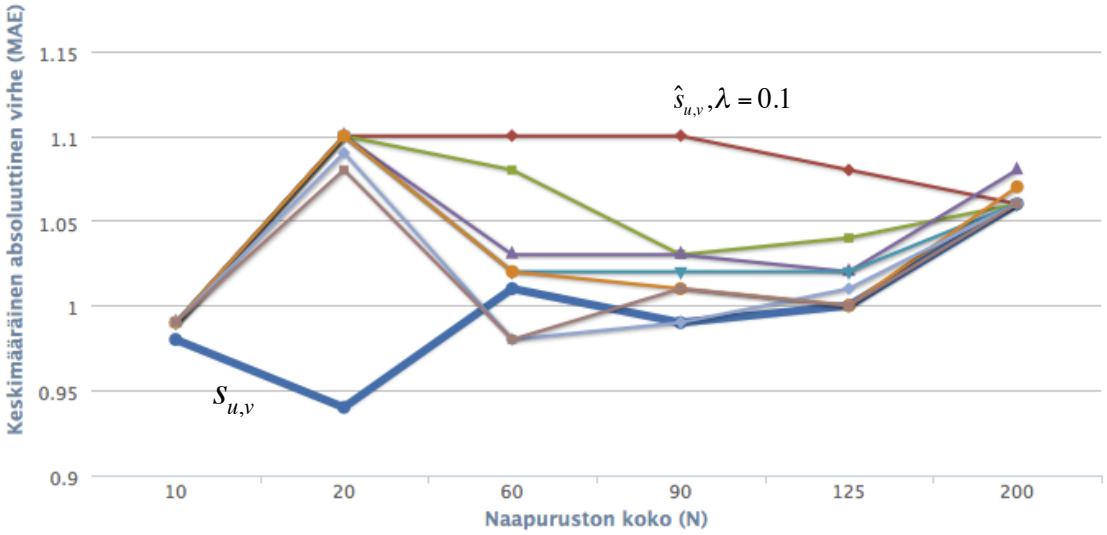
```
function predict($user, $movie, $users, $movies, $sim_f) {
    $nom_sum = 0; $dem_sum = 0;

    $avg = (float)db_get_value("SELECT AVG(rating) FROM
movie_rating WHERE uid = '$user' AND movie IN(" . join($movies,
',') . ")");

    foreach($users as $u) {
        $u_avg = (float)db_get_value("SELECT AVG(rating) FROM
movie_rating WHERE uid = '$u' AND movie IN(" . join($movies,
',') . ")");
        $sim = (float)db_get_value("SELECT $sim_f FROM
user_similarity WHERE uid1 = '$user' AND uid2 = '$u'");
        $rat = (float)db_get_value("SELECT rating FROM
movie_rating WHERE uid = '$u' AND movie = '$movie'");

        $nom_sum += $sim * ($rat - $u_avg);
        $dem_sum += $sim;
    }
    return $avg + $nom_sum / $dem_sum;
}
```

Lisäksi laskettiin absoluuttinen keskimääräinen virhe (MAE) määritelmän 2.6 mukaisesti. Virheen arvioinnissa pienin arvo tarkoittaa parhainta tarkkuutta. Kuvassa 4.7 on esitetty virhearvot tapauksille $\lambda = 0.9$ ja $\lambda = 0.8$ sekä puhdasta kosinietäisyyttä käyttäen. Näillä arvoilla suosittelutulokset paranevat jonkin verran pelkkään kosinietäisyyteen verrattuna – etenkin pienellä naapurustolla. Paras tulos saatiin arvolla $\lambda = 0.9$, jossa siis arvioista lasketun samankaltaisuuden paino on 90% ja sosiaalisen suhteen 10%. Tällä arvolla virhe oli jokaisella naapuruston koolla pienempi tai vähintään yhtä suuri kuin kosinietäisyydellä. Parhaimmillaan tällä arvolla prosentuaalinen parannus alkuperäiseen oli hieman yli 4% naapuruston koolla 20 ja hieman alle 5% naapuruston koolla 60. Kuvan 4.7 pystyakseli on katkaistu selvyyden vuoksi, jotta eri käyrät erottuisivat paremmin.



Kuva 4.7. *Sosiaalisen verkoston vaikutus suosittelutuloksiin, kun $\lambda = 0.9$ ja $\lambda = 0.8$.*

Kuvassa 4.8 on esitetty virhearvot tapauksille $\lambda = 0.1 \ldots 0.7$. Tämä tarkoittaa siis tapauksia, joissa sosiaalista verkostoa on painotettu 30-90%. Näillä arvoilla parannusta ei juurikaan ole. Naapurustoilla 20…125 tulokset ovat itse asiassa huonompia kuin oletustoteutuksella. Huonoin tulos saatiin arvolla $\lambda = 0.1$, eli lasketun samankaltaisuuden painoarvo on 10% ja sosiaalisen verkoston 90%. Kuvassa pystyakseli on katkaistu selvyyden vuoksi, jotta eri käyrät erottuisivat paremmin.



Kuva 4.8. *Sosiaalisen verkoston vaikutus suosittelutuloksiin, kun λ = 0.1…0.7.*

Vaikka tulokset ovat sinänsä hyvin subjektiivisia eivätkä välttämättä ole vertailukelpoisia, on niissä silti nähtävissä parannuksia suosittelutuloksiin. Tulokset ovat myös samansuuntaisia kuin Yuanin et al. (2009) tekemässä tutkimuksessa, jossa parhaat tulokset saatiin painottamalla sosiaalista suhdetta 20-30% pienellä naapurustolla.

5 JOHTOPÄÄTÖKSET JA POHDINTA

Tutkielman tarkoituksena oli selvittää, miten tietoa sosiaalisista verkostoista ja niiden suhteista voidaan hyödyntää suosittelujärjestelmässä. Päätutkimuskysymys, johon tutkielmassa haettiin vastausta on:

> *Onko mahdollista hyödyntää käyttäjän sosiaalista verkostoa suosittelujärjestelmässä?*

Taustalla on päähypoteesi, jonka mukaan tutuilta ihmisiltä saadut suositukset ovat parempia kuin tuntemattomilta käyttäjiltä tulevat. Hypoteesia testattiin simuloimalla yhteisöllistä suodatusta käyttävää järjestelmää, joka suosii käyttäjän samankaltaisten naapurien valinnassa sosiaalisen verkoston kautta tuttuja käyttäjiä. Simulointi tapahtui yhdistämällä olemassa olevaa dataa yhteisöllistä suodatusta käyttävästä suosittelujärjestelmästä kirjoittajan Facebook-kaveripiirille suunnatulla kyselytutkimuksella saatuun arvosteludataan.

Kyselyn kohteena ollut Facebook-kaveripiiri koostui kirjoitushetkellä 161 kontaktista. Näistä kontakteista vastauksensa kyselyyn antoi 21 henkilöä. Kyselyn kattavuus oli siis noin 13% kirjoittajan kaikista kontakteista. Kattavuus ei ole kovin suuri, mutta jos ajatellaan tilannetta oikean verkkopalvelun näkökulmasta on hyvin mahdollista, että vain pieni osa Facebook-kontakteista käyttäisi jotain tiettyä ulkoista sivustoa, jossa Facebookin verkostoa hyödynnetään.

Kyselyyn vastanneista henkilöistä arviolta viiden maku vastaa kirjoittajan omaa makua. Nämä henkilöt ovat myös sellaisia, joiden kanssa kirjoittaja usein tapaa sosiaalisesti. Kyselyyn vastanneista henkilöistä 12 oli ulkomaalaisia, jotka ovat kirjoittajalle tuttuja vaihto-opiskelujaksolta Unkarissa. Näiden henkilöiden kanssa yhteydenpito tapahtuu nykyään lähes kokonaan Facebookissa. Loput vastaajista olivat suomalaisia. Kuten kuvassa 4.6 on esitetty, käyttämällä pelkästään laskettua korrelaatiota, oli 20 samankaltaisimman käyttäjän joukossa vain kaksi Facebook-kontaktia.

Yli puolet kyselyyn vastanneista henkilöistä on kuitenkin sellaisia, joilta kirjoittaja uskoo saavansa mielenkiintoisia tai odottamattomia suosituksia, vaikka maut eivät suoraan korreloisikaan. Painottamalla sosiaalista suhdetta naapureiden valinnassa tulee mukaan luonnollisesti enemmän sosiaalisen verkoston kontakteja. Koska kyselyyn valittujen elokuvien lista oli varsin suppea, ei suosittelussa noussut esiin erityisen arvokkaita suosituksia. Suositusten laadun parantuminen luvussa 2.5.2 mainituin

kriteerein vaatisi laajemman käyttäjätutkimuksen, jossa esimerkiksi suositusten uutuusarvoa ja odottamattomia löytöjä voitaisiin testata.

Numeerisesti arvioiden simuloinnin tulokset olivat onnistuneita. Parhaimmassa tapauksessa, eli painottamalla sosiaalista verkostoa 10% ja valitsemalla naapuruston kooksi 20 tai 60, saatiin alkuperäiseen suositukseen 4-5% parannus. Vaikka tulos ei suoraan ole vertailukelpoinen, on kuitenkin hyvä huomioida, että Netflix maksoi miljoona dollaria tutkimusryhmälle, joka paransi heidän suosittelujärjestelmäänsä tarkkuutta kymmenellä prosentilla (Netflix 2012).

Vastauksena tutkimuskysymykseen voidaan sanoa, että sosiaalista verkostoa hyödyntämällä on mahdollista parantaa suosittelutuloksia. Pelkästään jo tässä tutkimuksessa suoritetulla yksinkertaisen sosiaalisen suhteen painotuksella saatiin aikaan havaittavia parannuksia. Ottamalla huomioon käyttäjien välisen luottamuksen tai muun sosiaalisen suhdearvon, voidaan mahdollisesti saada aikaan parempiakin tuloksia. Yleispätevät vastaukset vaativat kuitenkin vielä lisätutkimuksia.

Tutkielmassa keskityttiin vain elokuvien suositteluun ja sosiaalisen verkoston osalta ainoastaan Facebookiin, joten olisi mielenkiintoista tutkia millaisia tuloksia jollain toisella sosiaalisella verkostolla saataisiin. Facebookin sosiaalisen graafin ominaisuuksia ja sen antamia mahdollisuuksia ei myöskään täysimääräisesti hyödynnetty.

Seuraavissa aliluvuissa vastataan pääkysymyksen tarkentaviin kysymyksiin ja esitetään jatkokehitysideoita menetelmän käytölle.

5.1 Toteutuksen helppous

Ensimmäisessä tarkentavassa tutkimuskysymyksessä pyrittiin selvittämään suosittelujärjestelmän vaatimuksia ja edellytyksiä, joiden pohjalle voidaan rakentaa sosiaalista verkostoa hyödyntävä osa. Luonnollinen lähtökohta tähän on yhteisöllinen suodatus, jota käyttävä suosittelujärjestelmä koostuu suositeltavista tuotteista ja käyttäjistä, joihin sosiaalisen verkoston käyttäjät yhdistetään.

Kuten kuvassa 4.1 on esitetty, on sosiaalisen painotuksen lisääminen suosittelujärjestelmään varsin helppoa, sillä painotukseen käytetään jo olemassa olevia samankaltaisuusarvoja. Muut komponentit, eli samankaltaisuuden ja ennusteiden laskennan algoritmit, ovat sovelluskohtaisesti valittavissa. Menetelmä soveltuu kuitenkin vain muistipohjaiseen yhteisöllisen suodatukseen. Muihin suodatusmenetelmiin on olemassa vastaavanlaisia ratkaisuja. Esimerkiksi Yuan et al. (2009) tutkivat muistipohjaisen suodatuksen lisäksi myös sosiaalisten suhteiden yhdistämistä graafien avulla.

Muistipohjaisen suodatusmenetelmän heikkona puolena on laskennan vaatima aika. Tutkimuksessa samankaltaisuuden laskenta yhdelle käyttäjälle muiden, hieman alle 5700 käyttäjän suhteen kesti keskimäärin viisi sekuntia. Tietokantaan lisättäviä rivejä oli täten yhtä monta kuin käyttäjiäkin. Jos laskenta tehtäisiin 5700:lle käyttäjälle kaikkien käyttäjien suhteen, vaatisi laskenta aikaa 8 tuntia ja tietokanta tilaa noin 32

miljoonalle riville. Lisäksi laskenta pitää uusia joka kerta, kun tiedot muuttuvat. Skaalautuvuusongelmat ovat siis melkoiset. Parantamalla käytettyjä laskentatyökaluja ja laitteistoa ongelman laajuus pienenisi, mutta joka tapauksessa laskenta on niin työlästä, että sitä ei voida tehdä reaaliaikaisesti.

Sosiaalisen verkoston käyttöönotossa tärkein kriteeri on palveluiden tarjoamat rajapinnat. Tutkielmassa käytetty Facebookin Graph API -rajapinta tarjoaa kehittäjän saataville sosiaalisen graafin sekä muuta tietoa käyttäjän kiinnostuksen kohteista. Käyttämällä sosiaalisia liitännäisiä esimerkiksi palvelun kirjautumiseen, voidaan yhdistää kaikki suosittelujärjestelmän käyttäjät automaattisesti sosiaaliseen verkostoon. Tutkielmassa kokeiltu Facebook-integraatio oli varsin helppoa yksinkertaisen rajapinnan avulla. Tosin tässä tapauksessa kyseessä ei ollut todellinen verkkopalvelu vaan tiedot yhdistettiin käsityönä.

5.2 Suosittelutulosten paraneminen

Toista tarkentavaa tutkimuskysymystä varten suoritettiin kyselytutkimus. Kyselystä saatujen vastausten avulla mallinnetaan suosittelujärjestelmää, joka hyödyntää käyttäjän sosiaalista verkostoa. Viime vuosina tehdyissä tutkimuksissa (esimerkiksi Yuan et al. 2009 ja Golbeck 2006) saadut tulokset viittaavat siihen, että sosiaalisten suhteiden yhdistäminen perinteiseen yhteisölliseen suodatukseen on hyvä tapa tuoda lisätehokkuutta ja lisäarvoa suosittelujärjestelmään. Myös tässä tutkimuksessa kokeiltu sosiaalisten suhteiden ja yhteisöllisen suodatuksen yksinkertainen yhdistäminen näytti parantavan suosittelutuloksia.

Suosittelutuloksia arvioidessa tärkeää on silti sosiaalisen suhteen luotettavuus. Ihmiset kokevat arvokkaammiksi suositukset, jotka tulevat tutuilta henkilöiltä (Sinha & Swearingen 2001). Vaikka maut eivät suoraan korreloisikaan, on sosiaaliseen suhteeseen perustuvat suositukset todennäköisesti arvokkaampia kuin nimettömät elokuvapalvelun arviot vaikka suosittelijan tunteminen ei välttämättä takaakaan tuotteen sopivuutta omaan makuun (Lee & Brusilovsky 2009). Paljon tietysti riippuu myös käytetystä sosiaalisesta verkostosta.

Tässä tapauksessa kyselyn tutkimushenkilöt olivat Facebookista ja kaikki kyselyyn osallistuneet henkilöt olivat kirjoittajalle tuttuja myös sosiaalisesta kontekstista. Tällaisessa suhteessa palvelun ulkopuoliset mielipiteet tuovat enemmän painoarvoa käyttäjien tekemiin arvioihin tuotteista. Toisenlaista, esimerkiksi Last.fm-palvelusta saatua verkostoa käytettäessä tilanne ei välttämättä olisi niin latautunut.

Vaikka tulokset ovat sinänsä hyvin subjektiivisia, eivätkä ne välttämättä ole vertailukelpoisia, on niissä silti nähtävissä parannuksia suosittelutuloksiin. Tulokset ovat myös samansuuntaisia kuin Yuanin et al. (2009) tekemässä tutkimuksessa, jossa parhaat tulokset saatiin painottamalla sosiaalista suhdetta 20-30% pienellä naapurustolla. Nyt parhaat tulokset saatiin painottamalla sosiaalista suhdetta 10-20%. Parhaimmassa tapauksessa suositustulokset paranivat 4-5%.

Yhteisöllisen suodatuksen prosessin komponentit, kuten kuvassa 4.1 on esitetty, olivat tässä tutkimuksessa valittu yksinkertaisuuden perusteella. Parempia tuloksia voitaisiin tehdä optimoimalla sekä samankaltaisuuden että ennusteiden laskentaa. Tämä on kuitenkin sovelluskohtainen asia. Lisäksi tässä tutkimuksessa käytettiin sosiaalisen suhteen kuvaajana vain yksinkertaista yhteyttä, eli suhde joko on tai ei. Sosiaalisten suhteiden huomioon ottaminen näinkin yksinkertaisella tasolla näyttäisi kuitenkin olevan varteenotettava keino lisätä tarkkuutta suosittelutuloksiin. Tarkempia tai arvokkaampia suosituksia voisi olla mahdollista saada aikaan esimerkiksi tuomalla mukaan käyttäjien väliset luottamussuhteet tai ottamalla huomioon paremmin käyttäjän maun ja muut kiinnostuksen kohteet.

5.3 Suosittelujärjestelmien ongelmien kompensointi

Kolmannen tarkentavan tutkimuskysymyksen myötä pyritään pohtimaan keinoja, joilla voidaan estää tai kompensoida suosittelujärjestelmissä yleisesti havaittuja ongelmia.

Jos käyttäjällä on suosittelujärjestelmässä jo kontakteja, on sosiaalisen verkoston avulla mahdollista kompensoida uuden käyttäjän ongelmaa. Käyttäjälle, joka ei ole tehnyt yhtään arvostelua, on vaikea antaa osuvia suosituksia. Jos tiedetään kuitenkin, mitä käyttäjän kontaktit ovat arvostelleet, voidaan näitä tuotteita suositella aluksi. Tässä yhteydessä on myös hyvä esittää, kehen suositus perustui. Jos sosiaalisten suhteiden taustalla on jokin arvotus, kuten esimerkiksi luottamus, kannattaa näitä poimintoja luonnollisesti ottaa luotetuimmalta käyttäjältä. Sosiaalinen verkosto ei kuitenkaan auta tilanteessa, jossa käyttäjällä ei ole vielä kontakteja suosittelujärjestelmän sisällä.

Poikkeuksellisten yksilöiden eli käyttäjien, joiden maku on hyvin paljon valtavirrasta poikkeava, voi olla vaikeaa saada suosituksia automaattisin menetelmin. Tällainen käyttäjä voisi valita suoraan sosiaalisesta verkostostaan ne henkilöt, jotka hän tuntee ja joihin luottaa eniten suositteluissa. Esimerkiksi näyttämällä tuttujen käyttäjien nimi tai profiilikuva suosittelun yhteydessä (”myös kaverisi Jarno piti tästä tuotteesta”) voisi olla helppo keino lisätä mielenkiintoa uutta tuotetta kohtaan. Näyttämällä kaverit tai käyttäjät, joiden arvioista suodatus tehdään, auttaa järjestelmä käyttäjää ymmärtämään, mihin suosittelu perustuu.

5.4 Jatkokehitysideat

Vaikka tutkimus olikin hyvin subjektiivinen, vahvistaa se muun muassa Yuanin et al. (2009) havaitsemia tuloksia siitä, että suositustulokset paranevat hieman jo pelkällä tiedolla sosiaalisen suhteen olemassaolosta. On luonnollista ajatella, että ihmiset yleensä hakeutuvat toisten samankaltaisten ihmisten seuraan. Ei voida kuitenkaan sanoa, että kaikkien kavereiden maut olisivat yhteneväisiä.

Tutkimuksessaan Yuan et al. (2009) toteavat, että jäsenyys jossain ryhmässä (esimerkiksi ”Beatles”) vastaa paremmin suosittelun tarpeita kuin suora kaverisuhde. Tätä on hyödynnetty käyttämällä kolmatta painoarvoa sosiaalisen suhteen lisäksi

samankaltaisuuden laskennassa. Käyttämällä esimerkiksi sosiaalista graafia Last.fm-palvelusta, saattaisi tulos olla erilainen, sillä mukaan voidaan ottaa käyttäjän kuuluvuus artistien tai lajityyppien ryhmään. Toisaalta Facebook mallintaa paremmin jo olemassa olevia, reaalimaailman sosiaalisia suhteita. Last.fm-käyttäjät ovat ryhmittyneet lähinnä omien kiinnostusalueidensa ympärille.

Tutkielmassa hyödynnettiin Facebookin sosiaalista graafia vain käyttäjien yhteyksien osalta. Kaikki sosiaalisen graafin ominaisuuksia ei siis täysimääräisesti hyödynnetty. Facebook-käyttäjät voivat kertoa pitävänsä jostain sivusta tai kiinnostuksen kohteesta Facebookin tykkää-napin avulla. Tieto näistä on saatavilla sosiaalisen graafin kautta, joten tätä kannattaa hyödyntää menetelmän jatkokehityksessä esimerkiksi edellä mainitun kolmannen painoarvon tavoin.

Tutkimuksessa käytetty verkostodata ei sisällä mitään tietoa käyttäjien välisestä luottamussuhteesta. Jatkokehitysideana voisi olla luottamuksen liittäminen mukaan suosittelun laskentaan. Luvussa 3.4 mainittu FilmTrust-järjestelmä (Golbeck 2006) toimii kuten MovieLens-palvelukin, mutta käyttäjä määrää eksplisiittisesti, kuinka paljon kuhunkin käyttäjään luottaa. Heidän järjestelmässään luottamus määrätään asteikolla 1-10, jossa 1 on pienin ja 10 suurin mahdollinen luottamusta kuvaava luku. Skaala voi olla myös muunlainen.

Luottamuksen arvo ei välttämättä ole sama suhteen molemmin puolin. Esimerkiksi käyttäjän A luottamus käyttäjää B kohtaan voi olla arvoltaan 8, mutta käyttäjän B luottamus A:ta kohtaan vain 5. Hyödyntämällä eksplisiittistä luottamusarvoa voidaan tuottaa käyttäjälle entistä luotettavampia suosituksia. Huonona puolena on käyttäjälle koituva lisävaiva, koska hänen täytyy arvottaa luottamus kontakteihinsa. Tämä taas sopii huonosti ajatukseen automaattisesta suosittelujärjestelmästä. Eräs tapa toteuttaa tällainen ominaisuus automaattisissa järjestelmissä voisi kuitenkin olla jonkinlainen takaisinkytkentä tai palautenappula suositeltavan tuotteen yhteydessä.

6 YHTEENVETO

Tässä tutkielmassa esitettiin suosittelujärjestelmien toiminta yleisesti, keskittyen kuitenkin muistipohjaisiin suodatusmenetelmiin. Lisäksi esitettiin, miten sosiaaliset verkostoitumispalvelut ja sosiaaliset verkostot toimivat. Sosiaalisia verkostoja voidaan myös hyödyntää osana verkkopalvelun suosittelua. Tutkimuksen päätutkimuskysymyksenä oli, miten sosiaalinen verkosto toimii automaattisen suosittelun tukena. Lopuksi esitettiin tutkimus, jonka avulla kysymykseen haettiin vastausta.

Aiemmat tutkimukset ovat osoittaneet, että yhdistämällä sosiaalista verkostotietoa yhteisöllisen suodatuksen menetelmiin on mahdollista saada aikaan parannusta suosittelutuloksiin. Samoihin tuloksiin päädyttiin myös tässä tutkielmassa. Tarkoitus ei ollut parantaa olemassa olevia algoritmeja tai keksiä uusia, vaan kokeilla kuinka helppoa on ottaa olemassa olevaa dataa sosiaalisista verkostoista ja yhdistää ne olemassa olevaan suosittelutulokseen. Valitulla tutkimusaineistolla oli havaittavissa, että jo pelkän sosiaalisen suhteen mukaan ottaminen parantaa suosittelutuloksia hieman.

Tutkimuksen kohteena sosiaalisten verkostojen hyödyntäminen suosittelujärjestelmissä on varsin uusi. Menetelmän jatkokehittelyllä on kaupallistakin potentiaalia, sillä suosittelujärjestelmiä on käytössä lähes kaikissa verkkokaupoissa, muista palveluista puhumattakaan. Suosittelujärjestelmiltä vaaditaan koko ajan entistä parempia suosituksia. Lisäämällä mukaan tieto esimerkiksi eksplisiittisistä luottamussuhteista tai kiinnostuksen kohteista voitaisiin tuloksia parantaa mahdollisesti enemmänkin.

Tulevaisuudessa verkostojen saanti ja käsittely helpottunee vapaasti käytettävien rajapintojen avaamisen myötä. Jatkossa erilaiset sosiaaliset verkot on mahdollista koostaa osaksi suurempaa sosiaalista graafia, joka kokoaa käyttäjän kontaktit ja kiinnostukset yhden verkoston alle. Tällaisen verkoston käyttö suosittelussa mahdollistaa entistä parempia sovellutuksia.

LÄHTEET

Adomavicius, G. & Tuzhilin, A. 2005. *Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*. IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6. (25 June 2005), 734-749.

boyd, d. m. & Ellison, N. B. 2007. *Social Network Sites: Definition, History, and Scholarship*. Journal of Computer-Mediated Communication, Vol. 13, No. 1. (17 October 2007), 210-230.

boyd, d. m. & Heer, J. 2006. *Profiles as Conversation: Networked Identity Performance on Friendster*. Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS) (2006).

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A. & Riedl, J. 2003. *Is seeing believing?: how recommender system interfaces affect users' opinions*. CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems (2003), 585-592.

Dé, A. 2000. *Customer Fulfillment in the Digital Economy - Amazon.com: E-tail Customer Fulfillment Networks Pioneer*. 2000. Saatavilla http://www.andyde.com/amazon.pdf

Facebook. 2012a. *Fact Sheet*. Saatavilla: http://newsroom.fb.com/content/default.aspx?NewsAreaId=22

Facebook. 2012b. *Social Plugins*. Saatavilla https://developers.facebook.com/docs/plugins/

Fitzpatrick, B. 2007. *Thoughts on the Social Graph*. Saatavilla http://bradfitz.com/social-graph-problem/

Freeman, L. C. 2009. *Methods of Social Network Visualization*. Encyclopedia of Complexity and Systems Science. Berlin: Springer.

Golbeck, J. 2006. *Generating Predictive Movie Recommendations from Trust in Social Networks*. Trust Management, Vol. 3986 (2006), 93-104.

Golbeck, J. 2007. *The Dynamics of Web-based Social Networks: Membership, Relationships, and Change*. First Monday, 12(11), 2007.

Goldberg, D., Nichols, D. A., Oki, B. M. & Terry, D. B. 1992. *Using Collaborative Filtering to Weave an Information Tapestry*. Communications of the ACM, Vol. 35 (1992), 61-70.

Gong, S. 2010. *A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering.* Journal of Software, Vol. 5, No. 7. (01 July 2010).

GroupLens Research. 2011. *MovieLens Data Sets*. Saatavilla http://www.grouplens.org/node/73

GroupLens Research. 2012. *What is MovieLens?* Saatavilla http://movielens.org/html/tour/index.html

Gundotra, V. 2011. *Introducing the Google+ project: Real-life sharing, rethought for the web*. Saatavilla http://googleblog.blogspot.com/2011/06/introducing-google-project-real-life.html

Helin, T. & Koivisto, H. 2010. *Klusterointi*. Saatavilla http://www.ac.tut.fi/aci/courses/ACI-41050/Luennot/2010/Ryvastys2010_2p.pdf

Huang, J., Cheng, X., Guo, J., Shen, H., & Yang, K. *Social Recommendation with Interpersonal Influence.* Proceedings of ECAI (2010), 601-606.

JSON. 2012. *Introducing JSON*. Saatavilla http://www.json.org/

Kangas, P., Toivonen, S. & Bäck, A. 2007. *Googlen mainokset ja muita sosiaalisen median liiketoimintamalleja*. VTT Tiedotteita – Research Notes 2369. Espoo: VTT Information Technology, 2007.

Kangas, S. 2002. *Collaborative Filtering and Recommendation Systems*. Research Report TTE4-2001-35. Espoo: VTT Information Technology, 2002.

Kankainen, T. & Salminen, V-M. 2011. *Sosiaaliset verkostot*. Saatavilla http://kans.jyu.fi/sanasto/sanat-kansio/sosiaaliset-verkostot

Last.fm. 2012. *Frequently Asked Questions*. Saatavilla http://www.last.fm/help/faq

Lee, D. H. & Brusilovsky, P. 2009. *Does Trust Influence Information similarity?* Proceedings of the ACM RecSys'09 Workshop on Recommender Systems & the Social Web. New York, NY, USA.

Lietsala, K. & Sirkkunen, E. 2008. *Social media. Introduction to the tools and processes of participatory economy*. Hypermedia Laboratory Net Series 17. Tampere: Tampere University Press, 2008.

LinkedIn. 2012. *About Us*. 2012. Saatavilla http://press.linkedin.com/about

Marttila, J. 2010. *Datalähtöinen sosiaalisten verkostojen analyysi: tapaus Suomen Lasten Parlamentti*. Diplomityö. Tampere: Tampereen teknillinen yliopisto, 2010.

Milgram, S. 1967. *The Small World Problem*. Psychology Today (1967), Vol. 2, 60-67.

Mustonen-Ollila, E. 2006. *Relaatiotietokanta*. Saatavilla http://www2.it.lut.fi/kurssit/05-06/Ti5214300/kalvot.pdf

Netflix. 2012. *Netflix Prize*. Saatavilla http://www.netflixprize.com

Nichols, D. M. 1998. *Implicit rating and filtering*. Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering (1998), 31-36.

Pazzani, M. J. & Billsus, D. 2007. *Content-based Recommendation Systems.* The Adaptive Web, Vol. 4321 (2007), 325-341.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J. 1994. *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*. Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work (1994), 175-186.

Sanastokeskus TSK ry. 2010. *Sosiaalisen median sanasto*. Helsinki, Sanastokeskus TSK ry. Saatavilla http://www.tsk.fi/tiedostot/pdf/Sosiaalisen_median_sanasto

Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. 2001. *Item-Based Collaborative Filtering Recommendation Algorithms*. World Wide Web (2001), 285-295.

Schafer, J., Frankowski, D., Herlocker, J. & Sen, S. 2007. *Collaborative Filtering Recommender Systems*. The Adaptive Web, Vol. 4321 (2007), 291-324.

Sinha, R., & Swearingen, K. 2001. *Comparing recommendations made by online systems and friends*. Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries (2001).

Su, X. & Khoshgoftaar, T. M. 2009. *A Survey Of Collaborative Filtering Techniques*. Advances in Artificial Intelligence archive, Vol. 2009 (January 2009), 1-19.

Symeonidis, P., Nanopoulos, A., Papadopoulos, A. N. & Manolopoulos, Y. 2006. *Collaborative Filtering: Fallacies and Insights in Measuring Similarity*. Proceedings of the 10th PKDD Workshop on Web Mining (WEBMine'2006), 56-67, Berlin.

Tamada, S. 2011. *Facebook Graph API Connect with PHP and Jquery*. Saatavilla http://www.9lessons.info/2011/01/facebook-graph-api-connect-with-php-and.html

Tienvieri, V. 2010. *Verkostoitumispalvelut tieteellisessä tutkimuksessa.* Saatavilla http://verkkoaika.wordpress.com/2010/03/12/verkostoitumispalvelut-tieteellisessa-tutkimuksessa

Viljanen, K. 2006. *Monilähteinen suosittelu semanttisessa webissä*. Pro gradu -tutkielma. Helsinki: Helsingin yliopisto, 2006.

Wang, J., De Vries, A. & Reinders, M. 2006. *Unifying user-based and item-based collaborative filtering approaches by similarity fusion*. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (2006), 501-508.

Wasserman, S., & Faust, K. 1994. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Wei, D., Zhou, T., Cimini, G., Wu, P., Liu, W. & Zhang, Y-C. 2011. *Effective Mechanism for Social Recommendation of News*. Physica A, Volume 390, Issue 11, 2117-2126.

Wever, B., Mechant, P., Veevaete, P. & Hauttekeete, L. 2007. *E-Learning 2.0: Social Software for Educational Use*. Ninth IEEE International Symposium on Multimedia Workshops, 2007, ISMW '07, 511 – 516.

Yuan, Q., Zhao, S., Chen, L., Liu, Y., Ding, S., Zhang, X. & Zheng, W. *Augmenting Collaborative Recommenders by Fusing Social Relationships: Membership and Friendship*. Recommender Systems for the Social Web, Springer, 2012.

Zandstra, M. 2001. *PHP Trainer Kit*. IT Press.

LIITTEET

LIITE 1. *Kyselyyn valitut elokuvat.*

	Keskimääräinen arvio			Arvostelujen lukumäärä	
Elokuva	**MovieLens**	**Kysely**	**IMDB**	**MovieLens**	**Kysely**
2001: A Space Odyssey	4.1	3.6	8.4	1716	17
A Fish Called Wanda	4	4.3	7.7	1326	15
Ace Ventura: When Nature Calls	2.5	2.9	5.6	389	20
Apollo 13	4.1	3.1	7.5	1251	15
Austin Powers: International Man of Mystery	3.7	3.6	7.1	1205	14
Back to the Future	4	3.7	8.4	2583	19
Bad Boys	3.3	3.6	6.6	362	20
Batman Forever	2.6	2.9	5.4	777	14
Clueless	3.6	3	6.7	1362	19
Conan the Barbarian	3.4	3.3	6.8	572	9
Dances with Wolves	3.9	3.3	8	1451	18
Dead Poets Society	4	3.2	7.8	855	12
Desperado	3.5	2.7	7.1	540	9
Die Hard: With a Vengeance	3.6	3.3	7.5	825	15
Dumb & Dumber	3.2	3.1	7.2	660	15
Ed Wood	3.7	3	8.1	927	14
Face/Off	3.4	2.2	7.3	1421	15
Fargo	4.3	3.8	8.3	2513	12
Forrest Gump	4.1	4.4	8.6	2194	21
Four Weddings and a Funeral	3.7	3.3	7.1	1233	16
Goodfellas	4.3	3	8.8	1657	6
Groundhog Day	4	3.9	8.2	2278	14
Home Alone	3.1	3.1	7	675	14
Jackie Brown	3.7	3	7.6	724	16
Jerry Maguire	3.8	3.8	7.3	1353	18
Jurassic Park	3.8	3.2	7.9	2672	13
Liar Liar	3.5	2.4	6.7	666	18
Men in Black	3.7	3.7	7	2538	21
Monty Python and the Holy Grail	4.3	4.2	8.4	1599	20
Mrs. Doubtfire	3.4	2.8	6.6	838	13
Reservoir Dogs	4.1	4.1	8.4	1259	17
Scream	3.6	3	7.2	886	9
Sense and Sensibility	4	3.2	7.7	835	13
Starship Troopers	3.1	3.4	7.1	1163	22
The Big Lebowski	3.7	3.6	8.2	1097	15
The Blues Brothers	3.9	4.1	7.9	1341	11
The Bridges of Madison County	3.2	2.6	7.3	387	9
The Crying Game	3.7	3.2	7.3	1229	20

	Keskimääräinen arvio			Arvostelujen lukumäärä	
Elokuva	**MovieLens**	**Kysely**	**IMDB**	**MovieLens**	**Kysely**
The English Patient	3.6	3.8	7.3	989	13
The Fifth Element	3.6	2.8	7.5	1377	19
The Good, the Bad and the Ugly	4.1	4.1	9	822	14
The Lion King	3.9	3.8	8.2	1121	20
The Matrix	4.3	4.2	8.7	2590	20
The Silence of the Lambs	4.4	3.7	8.7	2578	17
The Truman Show	3.9	3.7	8	1005	17
Titanic	3.6	2.9	7.4	1546	22
Toy Story	4.1	2.9	8.2	2077	21
Trainspotting	4	3.8	8.2	751	20
True Lies	3.6	2.5	7.2	1400	8