hyväksymispäivä arvosana

arvostelija

Aine

Ola Länsman

Helsinki 14.4.2017

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

1 Johdanto

Pieni maailma -ilmiöksi kutsumme havaintoa lyhyiden ketjujen syntymisestä verkoissa ilman tietoa koko verkon kaarista. Verkkoja, joissa kyseinen ilmiö esiintyy, kutsutaan pieni maailma -verkoiksi ja näille verkoille on tehty useita malleja. Malliverkkojen luomisessa käytetään usein hyväksi parametreihin sidottua satunnaisuutta. Tälläisilla verkoilla on useita mielenkiintoisia ominaisuuksia, joita voidaan käyttää hyödyksi käytännön sovelluksissa. Suurin käytännön hyöty on lyhyiden polkujen etsiminen solmulta toiselle. Tätä varten on kehitetty useita algoritmejä, jotka pyrkivät optimoimaan polun pituuden ja polunetsintään kuluneen ajan.

Tutkielma selittää pieni maailma -ilmiön tieteellisesti alkaen tarkasta määrittelystä. Määrittelyn jälkeen näytämme, kuinka ilmiötä kuvaavia malleja voidaan luoda. Käytämme tähän käytännön esimerkkiä jota laajennamme yleisemmäksi muutamasta eri näkökulmasta katsoen. Luvussa 3 esittelemme polunetsintä-strategioita ja näihin perustuvia algoritmejä. Lyhyiden polkujen etsiminen solmujen välille on ilmiön tärkein sovelluskohde tietojenkäsittelytieteessä. Tästä pääsemme loogisesti käytännön sovelluksiin kuten vertaisverkossa resurssien etsimiseen ja automaattiseen luonnollisen kielen tiivistämiseen.

2 Pieni maailma

2.1 Määritelmä

Sosiaalisissa verkoissa pieni maailma -ilmiöksi kutstuaan havaintoa lyhyiden tuttavuusketjujen muodostuminen kahden eri yksilön välillä suurella todennäköisyydellä. Traversin ja Milgramin suurta huomiota saaeessa käytännön kokeessa [TM69] Yhdysvalloissa valittiin n. 200 lähettäjää ja vastaanottajaa. Lähettäjien tehtävänä oli lähettää viesti vastaanottajalle niin ,että viesti kulki ihmiseltä toiselle. Rajoitteena kokeessa viestin hallussapitäjä sai välittää viestin vain ihmiselle, jonka hän tunsi etunimellä. Kokeen tuloksena saatiin keskimäärin 6 pituisia tuttavuusketjuja, joka tunnetaan yleisesti "kuuden asteen erotuksena".

Verkkoa, jossa pieni maailma -ilmiö toteutuu, kutsumme *pieni maailma -verkoksi.* Tällaisia verkkoja ovat esimerkiksi erään madon hermoston muodostama verkko sekä verkko näyttelijöiden yhteistyön elokuvissa esiintymisestä[WS98]. Pieni maailma -verkoille ei ole olennaista kaarien suunta eikä esimerkiksi ruudukkomaisten verkkojen ulottuvuus. Niiden on kuitenkin oltava vahvasti yhtenäisiä, eli verkon jokaisen solmun on oltava saavutettavissa kaikista muista solmuista. Tietojenkäsittelytieteen verkkoteoriassa pieni maailma-verkoissa *suurella todennäköisyydellä minkä tahansa kahden solmun välille voidaan muodostaa lyhyt polku.* Polun lyhyys todetaan sen suhteesta verkon halkaisijaan. Verkon halkaisija on pisin lyhin polku kahden solmun välillä.

Pelkästään näillä ehdoilla pieni maailma -verkko ei ole välttämättä kovin mielenkiin-

toinen. Mikäli tekisimme verkon, jossa jokaisen solmupärin välille muodostuisi kaari jollain todennäköisyydelä p, suurella todennäköisyydellä kahden solmun välille löytyisi lyhyt polku parametrin p ollessa tarpeeksi suuri. Tällainen malli ei selitä, miten Traversin ja Milgramin tekemissä kokeissa viestin lähettäjät ovat kyenneet muodostamaan lyhyitä ketjuja käyttäen hyväksi ainoastaan viestin hallussapitäjän tietoja [Kle00]. Kutsummekin pieni maailma -verkkoa *navigoitavaksi*, mikäli *verkon yksittäiset solmut voivat lähettää viestin verkon kaaria pitkin muihin verkon solmuihin lyhyitä polkuja pitkin käyttämällä ainoastaan paikallista tietoa*. Paikallisella tiedolla tarkoitetaan, että viestiä hallussa pitävällä solmulla on tiedossa ainoastaan solmusta lähtevät kaaret.

Pieni maailma -verkkojen ominaisuuksia voidaan kuvailla Wattzin ja Strogatzin [WS98] määrittelemillä suureilla *keskimääräinen polun pituus* ja *ryhmittymiskerroin*. Keskimääräinen polun pituus verkossa kuvaa nimensä mukaisesti keskiarvoa kaikkien solmujen välisistä lyhimmistä poluista. Ryhmittymiskerroin $\mathcal{C}$ taas kuvaa kuinka hyvin lähekkäiset solmut ovat yhdistyneet toisiinsa. Esimerkiksi sosiaalisissa verkoissa suure vastaa kysymykseen, kuinka moni ystävistäni on ystävyksiä keskenään.

Määritelmä 1 (Ryhmittymisekerroin) *Olkoon $\mathcal{C}$ verkon G ryhmittymiskerroin, v suuntamattoman verkon G solmu ja solmulla v muuttujan k_v määrä naapureita. Tällöin solmun v ja sen naapurien välillä voi olla enintään $k_v(k_v - 1)/2$ kaarta. Olkoon $\mathcal{C}_v$ verkossa G oleva määrä kyseisistä sallituista kaarista. Määritämme ryhmittymiskertoimen $\mathcal{C}$ muuttujan $\mathcal{C}_v$ keskiarvoksi yli kaikkien solmujen v verkossa G.*

Verkon solmujen välille täytyy voida laskea paino jollain painofunktiolla, jolloin verkko pysyy navigoitavana. Yleisenä logiikkana voimme tällöin pitää, että viestiä lähettävällä on solmulla on jonkinlainen käsitys mihin suuntaan viestiä tulee lähettää, jotta se saavuttaa kohteen. Ihmiset tuntevat hyvin luultavasti toisensa paremmin etunimellä, jos asuvat lähekkäin. Jos tehtävänäsi olisi lähettää viesti sinulle tuntemattomalle henkilölle Lapissa, mahdollisesti lähettäisit viestin pohjoisessa asuvalle tutullesi. Tällöin maantieteellinen etäisyys muodostaa painon kaarille sosiaalisten suhteiden verkossa.

Navigoitavat pieni maailma-verkot yleensä täyttävät myös seuraavat ehdot [BBS11]:

1. *ne ovat harvoja*: Solmujen välillä on triviaalisti lyhyitä polkuja, jos verkon solmuilla on liikaa kaaria. Tällaiset verkot eivät ole mielenkiintoisia.

2. *ne ovat ryhmittyneitä*: Intuitiivisesti ehto 2 tarkoittaa, että mikäli solmu u ja v ovat lähellä toisiaan niin niiden välillä on todennäköisesti kaari. Yleisimmissä (ahneissa) polunetsintä-algoritmeissa viesti lähetetään aina mahdollisimman lähelle kohdetta ehdosta 2 johtuen.

3. *niillä on pieni halkaisija.*

Ehdossa 2 mainittulla solmujen läheisyydellä tarkoitetaan niiden välisen painon olevan pieni. Samaan ehtoon liittyy myös verkon *transitiivisuus*. Matemaattisesti yhteys R on transitiivinen, jos aRb ja bRc aiheuttavat suoraan aRc. Verkoissa transitiivisuutta voidaan ajatella seuraavalla tavalla: jos solmusta u on kaari solmuun v, ja solmusta v on kaari solmuun w niin solmulla u on kaari solmuun w. Vain täydelliset verkot voivat olla kokonaan transitiivisia. Transitiivisuus on myös yksi ryhmittymisen suure: mitä suurempi transitiivisuus, sitä ryhmittyneempi verkko on [BBS11].

Verkon transitiivisuus voidaan laskea yhdistyneiden kolmikoiden ja kolmioiden suhteena

$$C = \frac{(\text{kolmiot}) \times 3}{(\text{yhdistyneet kolmikot})}.$$

Kolmio tarkoittaa kolmen solmun joukkoja, joista jokaisen välillä on kaari. Yhdistynyt kolmikko tarkoittaa solmuja u, v ja w jossa solmusta u on kaari solmuihin v ja w. Osoittajan kerroin 3 tulee kolmioiden ja yhdistyneiden kolmikoiden laskutavasta. Yhdistyneitä kolmikoita laskiessa jokainen kolmio lasketaan kolmesti.

Algoritmejä, jotka käyttävät vain paikallista tietoa, kutsumme *hajautetuiksi algoritmeiksi*. Pieni maailma -verkkoja tarkastellessa tutkitaan monesti hajautettuja algoritmeja. Mikäli on mahdollista löytää hajautettu algoritmi, joka kykenee suorittamaan toimintonsa todennäköisesti tietyssä ajassa, pieni maailma -verkko on mielenkiintoinen. Algoritmin tehokkuutta kuvataan $\mathcal{O}$-notaatiolla ajan ja algoritmin *hyppyjen* suhteen. Algoritmin hypyillä tarkoitetaan, kuinka monella solmulla viesti on käynyt ennen kohteen saavuttamista. Tätä voidaan kutsua myös asteeksi.

2.2 Kuinka muodostetaan pieni maailma

Pieni maailma -verkkoja voidaan muodostaa monella tavalla. Aloitamme muodostamalla mallin, jonka kehitti Jon Kleinberg [Kle00] Wattz and Strogatz[WS98] -mallin pohjalta. Inspiraationa toimi Traversin ja Milgramin suorittama käytännön koe.

Mallinnamme ihmisiä solmuina. Tämä joukko solmuja muodostaa $n \times n$ ruudukon, missä

$$V = \{(i, j) : i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, n\}\}.$$

Ruudukkoetäisyytenä toimii Manhattan-etäisyys, eli

$$d((i, j), (k, l)) = |k - i| + |l - j|.$$

Solmulla u on *paikallinen kontakti* solmun v kanssa, jos $d(u, v) \leq p$ jollain vakiolla $p \geq 1$. Solmulla u on myös vakiomäärä, $q \geq 0$, *etäkontakteja*. Etäkontaktit solmujen u ja v välille muodostetaan satunnaisesti todennäköisyysfunktiolla, joka riippuu vakiosta $r \geq 0$ ja etäisyydestä $d(u, v)$. Tarkemmin, etäkontakti muodostetaan satunnaiskokeella solmusta u solmuun v todennäköisyydellä

$$\frac{[d(u, v)]^{-r}}{\Sigma_v [d(u, v)]^{-r}}$$

Edellä mainituilla tavoin muodostetut kaaret ovat suunnattuja. Solmun u etäkontaktilla v ei siis välttämättä ole kaarta takaisin solmuun u. Paikalliset kontaktit kuitenkin muodostavat suuntamattomia kaaria toistensa kanssa, koska niiden etäisyyteen liittyvä muodostamistapa on symmetrinen.

Tämä tapa muodostaa pieni maailma -verkko voidaan tulkita myös geometrisesti. Solmulla u on kaari jokaiseen tarpeeksi lähellä olevaan solmuun. Näiden yhteyksien lisäksi solmulla u on kaaria kauempana ruudukossa. Jos vakio $r = 0$, niin solmujen etäkontaktit ovat jakautuneet tasaisesti ruudukolle. Näin muodostettu verkko on pieni maailma, mutta se ei ole navigoitava. Vakion r kasvaessa solmun u etäkontaktit ovat jatkuvasti lähempänä solmua itseään.

Tässä mallissa ovat *etäkontaktit* mielenkiintoisimpia tarkastelun kohteita. Verkon navigoitavuuteen vaikuttaa, kuinka tasaisesti etäkontaktit ovat jakautuneet verkkoon. Jos $r = 0$, eli etäkontaktit olisivat jakautuneet tasaisesti verkkoon, niin ahne polunetsintä-algoritmi ei tuottaisi lyhyitä polkuja luotettavasti. Vaikka algoritmi löytäisi solmulta x kaaren solmuun y joka on lähellä kohdetta z, niin solmun y todennäköisyys omata etäkontakti kohteeseen z ei olisi kasvanut. Vakion r ollessa liian suuri olisi hyppyjen määrä myös liian suuri. Silloin viestit eivät pääsisi kulkemaan tarpeeksi pitkälle etäkontaktienkaan avulla.

Tarkemman tarkastelun jälkeen voimme huomata, että etäkontaktien ei tarvitse olla satunnaisesti tuotettuja luodaksemme navigoitava PM-verkko. Etäkontaktien satunnaisuutta vähentämällä verkosta muodostuu ryhmittyneempi, joka edesauttaa mm. vertaisverkkojen virheenkestävyyttä [CG06].

Rajoitamme etäkontaktien valitsemisen satunnaisuutta rajoittamalla solmun $u = (u_1, u_2)$ mahdollisiksi etäkontakteiksi vain solmut $v = (v_1, v_2)$, joille $u_1 = v_1$ tai $u_2 = v_2$. Tällöin solmun etäkontatkit sijaitsevat samalla suoralla solmun itsensä kanssa ja etäkontaktien valitsemisen satunnaisuus pienenee. Artikkelissa (inproceeding?) [CG06] on todistettu, että etäkontaktien muodostamista rajoittamisesta huolimatta verkon navigoitavuus säilyy. Edellisen ehdon lisäksi voitaisiin myös määritellä muita ehtoja, jotka koskevat vain osaa solmuista (*yhteisöt*), säilyttäen verkon pieni maailma -ominaisuudet.

3 Polunetsintä

Tässä luvussa esittelemme polunetsintä-algoritmejä, jotka soveltuvat pieni maailma -verkkoihin. Kutsumme tätä myös *reititykseksi*. Ensimmäisenä tutustumme yleisiin strategioihin, jonka jälkeen siirrymme varsinaisiin algoritmeihin. Esitämme algoritmeille aikavaativuuksien ylä- ja alarajoja ja esitämme myös ylärajoja algoritmien laatimien polkujen pituuksille.

3.1 Ahne reititys

Aiemmin esitetyssä käytännön kokeessa ihmiset muodostivat tuttavuusketjuja lähettäessään viestin kohteeseen. Voimme olettaa, että viestin lähettäjät pyrkivät lähettämään viestin tutulleen, jolla oli suurin todennäköisyys joko tuntea kohde tai kohteen tuttuja. Yksi hyvä strategia on viestin lähettäminen mahdollisimman lähelle kohdetta. Tämä on hyvä esimerkki *ahneesta strategiasta* jonka esittelemme seuraavaksi.

Ensimmäisenä tutkimme normaalia ahnetta strategiaa lyhyen polun etsintään. Ahneissa polunetsintäalgoritmeissa viestiä kuljettava solmu lähettää viestin aina lähimpänä kohdetta olevalle naapurilleen. Otamme esimerkiksi Kleinbergin [Kle00] esittämän termin *hajautettu algoritmi*, jossa viestiä kuljettava solmu tietää

1. kaikkien solmujen paikalliset kontaktit
2. kohteen y sijainti ruudukossa
3. kaikkien viestiä kuljettaneiden solmujen etäkontaktit ja sijainnit.

Osoitamme, että eräälle luvussa 2.2 esitetyllä mallilla muodostetulle pieni maailma -verkolle voidaan kehittää hajautettu algoritmi, jonka keskimääräinen *hyppyjen* (monellako solmulla viesti on käynyt) määrä on $\mathcal{O}(log^2 n)$. Kyseisellä mallilla parametrien arvot ovat $r = 2$ ja $p = q = 1$. Näytämme myös, että tämä on ainoa Kleinbergin esittämistä malleista, josta voidaan muodostaa navigoituva pieni maailma -verkko jossa hajautetut algoritmit ovat tehokkaita. Huomaammekin, kuinka tärkeä verkon rakenne on lyhyiden reittien löytämisen vain paikallisella tiedolla.

Etäkontaktit levittyätyvät verkkoon tasaisesti, kun parametri $r = 0$. Tällöin minkä tahansa kahden solmun välille löytyy lyhyt reitti suurella todennäköisyydellä. Hajautettu algoritmimme ei kuitenkaan kykene löytämään sitä luotettavasti, sillä verkolla ei ole heuristiikkaa, jota algoritmi voisi hyödyntää.

Lause 1 *On olemassa vakio α_0, joka riippuu vakioista p ja q, niin että, kun vakio $r = 0$, on hyppyjen odotusarvo vähintään $\alpha_0 n^{2/3}$.*

Kun parametria r kasvatetaan, kykenee algoritmi hyödyntämään syntyvää heuristiikkaa. Tämä parametrin r arvo on 2.

Lause 2 *On olemassa hajautettu algoritmi $\mathcal{A}$ ja vakio α_2, riippumaton vakiosta n, niin että kun $r = 2$ ja $p = q = 1$, niin algoritmin $\mathcal{A}$ hyppyjen odotusarvo on enimmillään $\alpha_2 (\log n)$.*

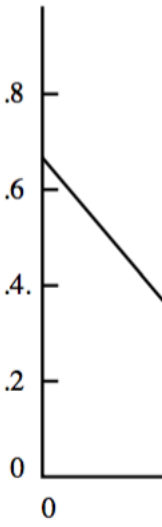
Lauseiden 1 ja 2 todistusten taustalla olevat ideat esitetään seuraavaksi. Lauseen 2 hajautettu algoritmi $\mathcal{A}$ etenee vaiheittain: sanommekin, että algoritmi on vaiheessa j, jos käsittelyssä olevan solmun etäisyys kohteeseen on arvojen j^2 ja j^{2+1} välissä.

Voidaan osoittaa, että odotusarvo hyppyjen määrälle ennen kuin käsittelyssä olevalla solmulla on etäkontakti solmuun, jonka etäisyys kohteesta on pienempi kuin j^2, on verrannollinen arvoon $\log n$. Tällöin odottettu hyppyjen määrä on enimmillään $\log n + 1$. Lauseen 1 raja perustuu todennäköisyyslaskentaan. Muodostamme joukon U, johon kuuluvat kaikki solmut joiden etäisyys kohteeseen on enimmillään $n^{2/3}$. Tällöin lähtösolmu suurella todennäköisyydellä ei kuulu joukkoon U, jolloin matka kohteeseen ilman etäkontakteja on ainakin $n^{2/3}$. Todennäköisyys sille, että millä tahansa solmulla olisi etäkontakti joukossa U on kuitenkin karkeasti suhteessa arvoon $n^{-2/3}$. Tämän seurauksena odotettujen hyppyjen määrä on suhteessa arvoon $n^{2/3}$.

Seuraavien lauseiden pohjalta voimme todeta, että odotettu hyppyjen määrä on logaritmisesti suhteessa solmujen määrään ainoastaan, kun $r = 0$.

Lause 3 *(a) Olkoon $0 \leq r \leq 2$. On olemassa vakio α_r, riippuen vakioista p, q, r, mutta riippumaton vakiosta n, niin että odotettu toimitusaika mille tahana hajautetulle algoritmille on ainakin $\alpha_r n^{(2-r)/3}$.*

(b) Olkoon $r > 2$. On olemassa vakio α_r, riippuen vakioista p, q, r, mutta riippumaton vakiosta n, niin että odotettu toimitusaika mille tahana hajautetulle algoritmille on ainakin $\alpha_r (r-2)/(r-1)$.



Kuva 1 (Odotettujen hyppyjen määrä suhteessa parametrin r arvoon.)

Lauseen 3 (a) todistus on samalainen kuin lauseen 1. Kohdan 3 (b) todistus sen sijaan perustuu etäkontaktien liialliseen keskittymiseen solmun lähelle. Algoritmi ei enää pääse tarpeeksi lähelle kohdetta etäkontaktien avullakaan.

3.2 Epäsuora reititys ja Naapurien naapurit

Epäsuora ahne reititys toimii kuten ahne reititys. Poikkeuksena, viestiä kuljettavalla solmulla u on tiedossa myös solmun v etäkontaktit jos solmulle pätee $d(u, v) \leq q$

jollain etäisyysfunktiolla d ja vakiolla q. Tällöin viestiä kuljettavalla solmulla on mahdollisuus lähettää viesti jonkin itseään lähellä olevan solmun kautta. Naapurien naapurit -strategia toimii samalla periaatteella kuin epäsuora ahne reititys. Lähellä olevien solmujen sijaan viestiä kuljettavalla solmulla on tiedossa omien naapureidensa kontaktit. Solmun naapureilla tarkoitamme solmuja, joilla on joko etäkontakti tai paikallinen kontakti kyseiseen solmuun.

Seuraavaksi esitämme erään Naapurien naapurit -algoritmin, jota kutsumme algoritmiksi $\mathcal{N}$. Kiinnitämme huomiota muodostetun polun pituuteen, jonka merkitys on suurempi mm. vertaisverkkosovelluksissa. Erityisesti vertaamme sitä luvussa 3.1 mainittuun algoritmiin. Tulokset eivät ole täysin verrannollisia, koska emme käytä lukuun 2.2 perustuvaa mallia molempien algoritmien tulosten toteuttamiseen. Poistamme mallista konseptit etä- ja lähikontaktit, mutta pidämme lauseessa 2 määritellyn parametrin r arvon 2 mukana. Muodostamme solmulle u kaaren solmuun v todennäköisyydellä $d(u, v)^{-2}$.

Algoritmin $\mathcal{N}$ viestin lähettämisestä solmuun u etenee seuraavasti:

1. Oletetaan viestin olevan solmussa $u \neq t$. Olkoon solmut $w_1, w_2, \ldots, w_k$, solmun u naapureita.
2. Kullekin i olkoon solmut $z_{i_1}, z_{i_3}, \ldots, z_{i_k}$ solmun w_i naapureita.
3. Oletetaan solmun z_{i_j} olevan tästä joukosta lähimpänä kohdetta t.
4. Lähetetään viesti solmuun z_{i_j} solmun u_j kautta.

Algoritmin $\mathcal{N}$ keskimääräinen hyppyjen määrä mallissamme suurella todennäköisyydellä on $\mathcal{O}(\log n / \log \log n)$ [MNW04]. Intuitiivisestikin voidaan ajatella, että naapurien naapurit -strategia löytää tehokkaammin lyhyitä reittejä kuin ahne reititys, jonka odotetty hyppyjen määrä oli *logn* samankaltaisessa verkossa. Tämä tulos ei kuitenkaan suoraan todista väitettä.

Epäsuoraan ahneeseen reititykseen ja naapurien naapurit -strategiaan liittyy kuitenkin haittapuolensa. Nämä ovat lisätietojen vaatima muistin käytön kasvu useampien solmujen naapurien säilyttämiseen. Käsittelemme näitä tarkemmin seuraavassa luvussa vertaisverkkojen ohella.

4 Käytännön sovellukset

Tässä luvussa esittelemme, kuinka pieni maailma -ilmiötä voidaan käyttää hyväksi pieni maailma -verkoissa. Kuten aiemmin sanottu, pieni maailma -ilmiötä löytyy monista biologista, sosiaalisista ja teknologista verkoista. Sovelluksia tarkastellessamme käytämme hyväksi lukujen 2 ja 3 tuloksia ja käsitteitä.

4.1 Peer-to-peer verkot

Peer-to-peer verkoissa, eli vertaisverkoissa, ovat asiakkaat ja palvelimet ovat sama asia. Jokainen verkkoon liittyvä kone voi käyttää muiden verkossa olevien koneiden resursseja, sekä sen resursseja voidaan käyttää. Nämä resurssit voivat olla esimerkiksi levytila tai laskentateho. Meitä kiinnostava kohde ovat juuri nämä resurssit ja niiden löydettävyys. Löydettävyys on tärkeä osa vertaisverkkoja, koska koneet eivät voi ylläpitää yhteyttä joka ainoaan muuhun verkon koneeseen. Tiettyjä palveluita joita voivat tarjota vain osa koneista on kuitenkin kyettävä löytämään tästä huolimatta aiheuttamatta liikaa kuormitusta. Pieni maailma -verkot toimivatkin motivaationa vertaisverkoille. Aiemmin esittämiemme tulosten perusteella, yhteyksiä muihin koneisiin ei tarvitse olla useita ja silti puuttuvat resurssit voidaan löytää tehokkaasti.

Vertaisverkkojen huonoja puolia ovat

4.2 Yhteenvedot luonnollisen kielen teksteistä

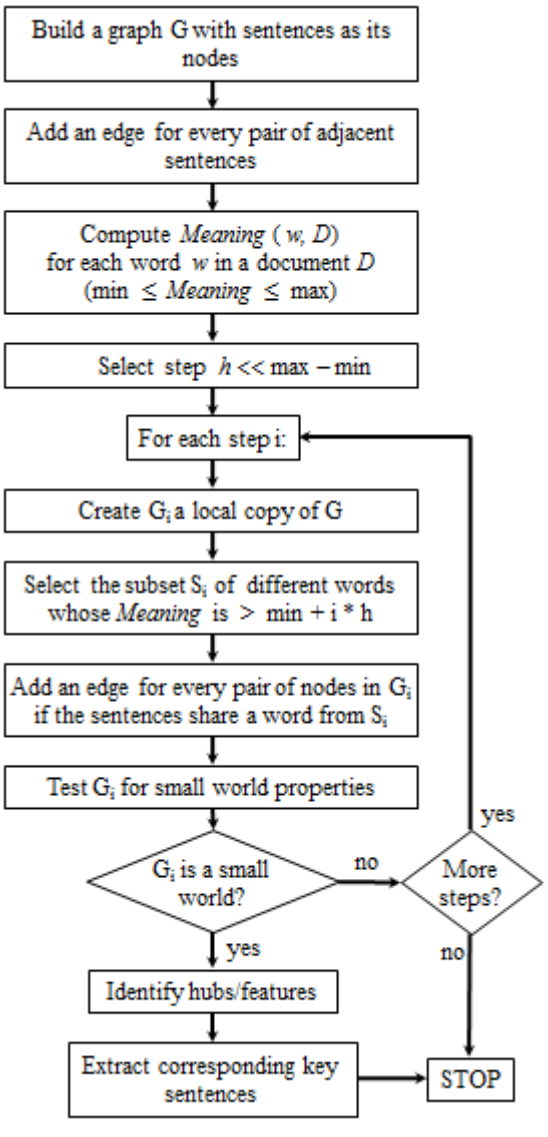
Luonnollisten kielten tekstien asiasisällön esittäminen tiiviimmin on hyödyllistä nykymaailmassa elektronisesti käsillä olevan tiedon kasvaessa räjähdysmäisesti. Tekstin sisällön voi esittää lyhyemmin muodostamalla uuden tekstin tai valikoimalla tekstistä tärkeimmät virkkeet ja kappaleet. Seuraavaksi esittelemme virkkeitä ja kappaleita valikoivan algoritmin, joka käyttää hyväkseen pieni maailma -ilmiötä [BBS11] . Algoritmi rakentaa tekstin virkkeistä pieni maailma -verkon ja poimii virkkeistä ne, jotka edesauttavat verkkoa eniten olemaan pieni maailma.

Aloitamme määrittelemällä verkon $G = (V, E)$, jossa pisteet V ovat virkkeitä ja kaaret E kuvaavat virkkeiden välisiä suhteita. Virkkeellä L on lähikontakti virkkeen L' kanssa, jos virkkeet L ja L' ovat peräkkäisiä. Etäkontaktien muodostamiseen tarvitsemme keinon määrittää virkkeiden yhteyden toisiinsa, joka muodostaa verkon heuristiikan. Tätä varten rakennamme tekstin tärkeimmistä sanoista joukon $MeaningfulSet(e)$. Etäkontakti kahden virkkeen välille muodostuu vain, jos kummassakin virkkeessä esiintyy jokin joukon $MeaningfulSet(e)$ sanoista. Parametri e määrittää, kuinka suuri joukko $MeaningfulSet(e)$ on, asettamalla rajan jota suurempi sanan *merkityksen* on oltava, jotta se valitaan joukkoon $MeaningfulSet(e)$. Sanan merkitys voidaan laskea jollekin osajoukolle täydestä tekstistä Helmholtz periaatteen avulla. Nämä osajoukot voivat olla esimerkiksi kappaleita tai viiden virkkeen joukkoja teksteissä, joissa kappalejakoa ei ole. Sanan merkitystä laskiessa käytetään hyväksi sanan esiintymisten lukumäärää koko tekstissä ja kappaleessa. Tällä tavoin konjunktioille ja muille yleisille sanoille saadaan pieni merkitys, mikäli sana esiintyy koko tekstissä erittäin monesti. Sanan merkitykseksi valitaan eri osajoukoille lasketuista merkityksistä suurin.

Joukon MeaningfulSet(e) sanojen määrä suhteessa joukkoon kaikista tekstissä esiintyvistä sanoista vaikuttaa verkon G kaarien määrän ja täten myös tiivistelmän pituuteen ja olennaisuuteen. Jos joukko $MeaningfulSet(e)$ on liian suuri, verkko G

ei näytä enää pieneltä maailmalta vaan sattumanvaraisesti muodostetulta verkolta. Kuitenkin joukon $MeaningfulSet(e)$ ollessa liian pieni verkko G näyttää molempiin suuntiin linkitetyltä listalta josta löytyy muutama poikkeus. Yhteenvetomenetelmän toimintaperiaatteen kannalta tällöin on suuresti merkitystä, miten tämä joukko valitaan, joten parametrin e valinta on tärkeää. Tätä ongelmaa voidaan verrata luvussa 2.2 esitettyyn pieni maailma -verkon malliin. Myös tiivistämisalgoritmia varten pieni maailma -verkon rakentamisen kannalta on parametrilla e tietty arvo, jolloin usein verkosta muodostuu tehokas pieni maailma.

Kuva 2 (Automaattinen tekstin tiivistäminen)



//

Algoritmin toimintaan liittyy oleellisesti myös pieni maailma -verkkojen ominaisuuksien testaaminen. Tähän algoritmi käyttää aiemmin määrittelemiämme suureita ryhmittymiskerroin ja transitiivisuus. Mikäli virkkeistä muodostettu verkko ei toteuta annettuja ehtoja ryhmittymiskertoimen ja transitiivisuuden suhteen, asetetaan parametrille e uusi arvo ja yritetään uudestaan. Ehtojen vihdoin täytyttyä

etsitään virkkeet jotka eniten edistävät verkkoa olemaan pieni maailma. Nämä virkkeet muodostavat yhteenvedon. Virkkeiden panosta pienen maailman rakentumiseen voidaan mitata tietyillä mittareilla. Näille mittareille on tärkeää suuri arvojoukko, jolloin yhteenveto saadaan muodostettua tarpeeksi lyhyeksi kaikkein merkityksellisimmistä virkkeistä.

Seuraava kappale on Yhdysvaltain presidentin vuonna 2011 pitämän Kansakunnan tila -puheen tärkein kappale parametrin e arvolla 2.

The plan that has made all of this possible, from the tax cuts to the jobs, is the Recovery Act. That's right, the Recovery Act, also known as the stimulus bill. Economists on the left and the right say this bill has helped save jobs and avert disaster. But you don't have to take their word for it. Talk to the small business in Phoenix that will triple its workforce because of the Recovery Act. Talk to the window manufacturer in Philadelphia who said he used to be skeptical about the Recovery Act, until he had to add two more work shifts just because of the business it created. Talk to the single teacher raising two kids who was told by her principal in the last week of school that because of the Recovery Act, she wouldn't be laid off after all.

Tärkein lause taas oli *The plan that has made all of this possible, from the tax cuts to the jobs, is the Recovery Act.* Nämä lainaukset johtavatkin suureen haasteeseen, nimittäin yhteenvetojen arviointiin. Olisiko pelkkä ensimmäinen kappale riittävä yhteenveto Obaman puheesta?

Lähteet

BBS11 Balinsky, H., Balinsky, A. ja Simske, S. J., Automatic text summarization and small-world networks. *Proceedings of the 11th ACM Symposium on Document Engineering*, DocEng '11, New York, NY, USA, 2011, ACM, sivut 175–184, URL `http://doi.acm.org/10.1145/2034691.2034731`.

CG06 Cordasco, G. ja Gargano, L., How much independent should individual contacts be to form a small–world? *Proceedings of the 17th International Conference on Algorithms and Computation*, ISAAC'06, Berlin, Heidelberg, 2006, Springer-Verlag, sivut 328–338, URL `http://dx.doi.org.libproxy.helsinki.fi/10.1007/11940128_34`.

DHLS06 Duchon, P., Hanusse, N., Lebhar, E. ja Schabanel, N., Could any graph be turned into a small-world? *Theoretical Computer Science*, 355,1(2006), sivut 96 – 103. URL `http://www.sciencedirect.com/science/article/pii/S0304397505009187`.

Kle00 Kleinberg, J., The small-world phenomenon: An algorithmic perspective. *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, New York, NY, USA, 2000, ACM, si-

vut 163–170, URL http://doi.acm.org.libproxy.helsinki.fi/10.1145/335305.335325.

MNW04 Manku, G. S., Naor, M. ja Wieder, U., Know thy neighbor's neighbor: The power of lookahead in randomized p2p networks. *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, New York, NY, USA, 2004, ACM, sivut 54–63, URL http://doi.acm.org.libproxy.helsinki.fi/10.1145/1007352.1007368.

TM69 Travers, J. ja Milgram, S., An experimental study of the small world problem. *Sociometry*, 32,4(1969), sivut 425–443. URL http://www.jstor.org/stable/2786545.

WS98 Watts, D. J. ja Strogatz, S. H., Collective dynamics of 'small-world'networks. *nature*, 393,6684(1998), sivut 440–442.