Anri Patron

Tekstin representointi katkaistulla pääakselihajotelmalla luokittelussa

Tietotekniikan kandidaatintutkielma

30. huhtikuuta 2019

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Anri Patron

Yhteystiedot: anri.patron@gmail.com

Ohjaaja: Antti-Jussi Lakanen

Työn nimi: Tekstin representointi katkaistulla pääakselihajotelmalla luokittelussa

Title in English: Representing text with truncated singular value decomposition in categorization

Työ: Kandidaatintutkielma

Sivumäärä: 25+0

Tiivistelmä: Tekstin representaatio on kiinteä osa luonnollisen kielen prosessointia, sillä se mahdollistaa luonnollisten kielten laskennallisen analysoinnin. Yleiset representaatiomenetelmät ovat syntaksiin perustuvia. Luonnolliseen kieleen liittyy kuitenkin olennaisesti tulkinnanvaraisuutta, mikä aiheuttaa syntaktisiin representaatioihin vääristymiä. Tutkielmassa tarkastellaan tekstin representointia katkaistulla pääakselihajotelmalla luokitteluongelman näkökulmasta. Pääakselihajotelmalla approksimoimalla tekstiaineistosta voidaan löytää termien ja dokumenttien assosiatiivisten yhteyksien rakenne, jota voidaan käyttää tekstin representointiin. Menetelmällä saatavat tulokset vaikuttavat lupaavilta syntaksiin perustuviin representaatiomentelmiin verrattuna.

Avainsanat: teksti, representaatio, pääakselihajotelma, luokittelu

Abstract: Text representation is a critical part of natural language processing and a prerequisite for any computational analysis. Popular representational methods are based on syntactic terms. However interpretability of natural language causes noise in syntactic representations. This paper evaluates the use of truncated singular value decomposition as text representation in text categorization. Singular value decomposition is used in transforming original term by document matrix into a subspace where text is represented as associations of terms and documents. Results show truncated singular value decomposition to be promising replacement for syntactic representation methods.

Keywords: text, representation, singular value decomposition, categorization

Kuviot

Sisältö

1 Johdanto

Luokittelulla (eng. categorization) tarkoitetaan objektien asettamista olemassa oleviin kategorioihin. Luokituksella (eng. classification) taas tarkoitetaan objektien luokittelua, kun kategorioita ei ole annettu. Tässä tutkielmassa käsitellään aiemmin mainittua luokittelua luonnollisen kielen prosessoinnin näkökulmasta. Tekstin luokittelua hyödynnetään esimerkiksi dokumenttien organisointiin, tekstin suodattamiseen ja tekstistä luettavien tunnetilojen analysointiin (eng. sentiment analysis).

Luonnollisen kielen prosessointi edellyttää dokumenttien tai tekstin mallintamista rakenteellisesti. Näistä mallinnetuista dokumenteista puhutaan dokumenttien representaatioina. Luonnollisen kielen prosessoinnissa suositut tekstin representaatiokeinot ovat syntaksipohjaisia, missä dokumentit representoidaan yksinkertaisesti niiden sisältämien termien joukkoina. Tyypillisessä representaatiossa termeiksi valitaan välilyönnin erottamat sanat. Yleensä syntaksipohjaiset representaatiokeinot eivät ota huomioon sanajärjestystä, joten samoista termeistä koostuvat dokumentit tulkittaisiin samoiksi. Vastaavasti kahden tai useamman sanan muodostamat kokonaisuudet hajoavat ja siten niiden merkitys voi muuttua tai hämärtyä, esimerkiksi sanapari ’New York’ hajoaisi termeiksi ’new’ ja ’york’. Tällaista representaatiomallia kutsutaan BoW-malliksi (Bag-of-Words -malliksi).

Luonnollisen kielen prosessoinnissa dokumenttien vertailu perustuu tyypillisesti termien esiintyvyyksien vertailuun. BoW-mallissa dokumenttien vertailu on haastavaa, koska ihmiset voivat kuvailla samaa informaatiota täysin eri termein. Esimerkki tästä ilmiöstä on synonyymiset sanat. Synonyymiset sanat voivat aiheuttaa todellisuudessa samankaltaisten dokumenttien vertailussa negatiivisen tuloksen. Tämän lisäksi luonnollisissa kielissä termien merkitys voi olla kontekstisidonnainen ja tulkinnanvarainen. Esimerkiksi termillä ’kuusi’ voidaan viitata numeroon tai havupuuhun. Nämä eri merkitykset voivat aiheuttaa todellisuudessa epärelevanttien dokumenttien tulkinnan samankaltaisiksi.

Yllä kuvattuja ongelmia on pyritty ratkaisemaan mallintamalla dokumentteja syntaktisten termien sijasta konseptuaalisilla piirteillä. Konseptilla viitataan termien taustalla olevaan merkitykseen. Esimerkiksi termit ’mänty’ ja ’koivu’ ovat eri termejä, mutta ne voivat vii-

tata samaa ylätason konseptiin. Pääakselihajotelma on eräs faktorianalyysin menetelmä, jonka avulla on mahdollista löytää tekstiaineistosta konseptuaalisia piirteitä, jotka kuvaavat termien ja dokumenttien välisiä yhteyksiä (Deerwester ym. 1990). Pääakselihajotelman avulla tekstiaineisto voidaan uudelleenparametrisoida, joka mahdollistaa dokumenttien vertailun konseptuaalisella tasolla. Uudelleenparametrisoinnin myötä yhteyksiä voidaan löytää myös sellaisten dokumenttien väliltä, jotka eivät sisällä samoja termejä. Tutkielmassa tarkastellaan, miten tekstiä voidaan representoida pääakselihajotelmalla approksimoimalla ja miten tämä representaatiomenetelmä soveltuu luokitteluongelmiin.

Tutkielman myöhemmät kappaleet on jäsennelty seuraavasti. Kappaleessa 2 esitellään luonnollisen kielen prosessoinnin taustatietoja, kuten yleisesti käytetty vektoriavaruusmalli, jota usein kutsutaan BoW-malliksi. Kappaleessa esitellään myös luonnollisen kielen prosessoinnissa keskeiset painofunktiot ja esikäsittelymenetelmät. Kappaleessa 3 määritellään pääakselihajotelma ja sen käyttö tekstin representointiin kuvaillaan. Kappaleessa 4 tarkastellaan tekstin representointia pääakselihajotelma-approksimaatiolla luokittelussa. Tutkielman viimeisessä kappaleessa 5 esitetään tutkielman yhteenveto ja jatkotutkimusehdotuksia.

2 Tekstin representointi

Luonnollisen kielen käsitteleminen laskennallisesti vaatii jonkinlaisen rakenteellisen representaation. Saltonin, Wongin ja Yangin (1975) esittelemä vektoriavaruusmalli on yksinkertainen tapa esittää tekstidataa rakenteellisessa muodossa. Vektoriavaruusmallissa dokumenttijoukon $D = (d_1, d_2, ..., d_n)$ jokainen dokumentti esitetään vektorina $d_i = (w_{i_1}, w_{i_2}, ..., w_{i_j})$ termiavaruudessa, jossa w_{i_j} on termin j paino dokumentissa i. Kuviossa 1 on graafisesti esitetty vektoriavaruusmalli, jossa dokumentit esitetään vektoreina kolmedimensioisessa avaruudessa.

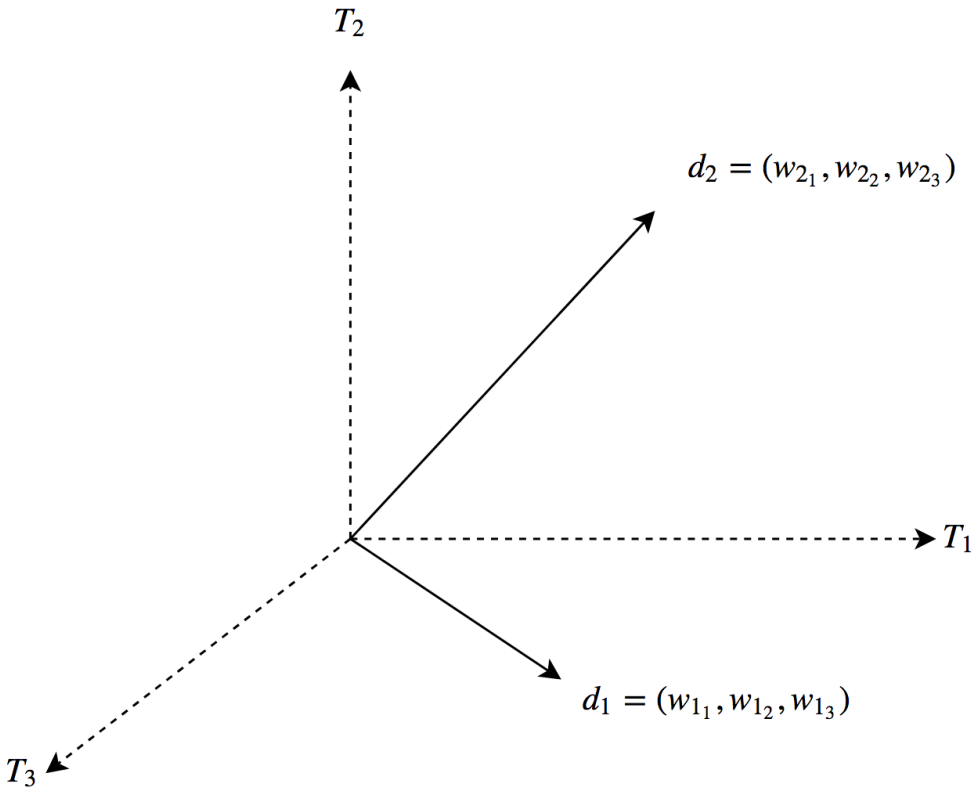
Kahden dokumentin samankaltaisuutta voidaan arvioida vektoriavaruusmallissa esimerkiksi mittaamalla dokumentteja vastaavien vektoreiden välisen kulman suuruutta. Eräs suosittu tapa laskea dokumenttivektoreiden samankaltaisuuskerroin on kosinisimilaarisuusfunktio, joka on rajoitettu $x \in [0, 1]$, kun painot ovat positiivisia (Salton ja Buckley 1988). Kosinisimilaarisuusfunktio määritellään seuraavasti

$$sim(d_a, d_b) = \frac{d_a \cdot d_b}{\|d_a\|\|d_b\|} = \frac{\sum_{i=1}^{n} w_{a_i} w_{b_i}}{\sqrt{\sum_{i=1}^{n} (w_{a_i})^2 \sum_{i=1}^{n} (w_{b_i})^2}}. \tag{2.1}$$

2.1 Painotus

Kaikki termit eivät ole yhtä tärkeitä dokumentin representaation kannalta, sillä osa termeistä kuvaa dokumentin sisältöä paremmin kuin toiset. Termien painotuksella voidaan määritellä termien relevanttius dokumenteille. Kenties yksinkertaisin painotusskeema on painottaa termejä niiden frekvenssin, eli esiintyvyyden mukaan. Luhnin (1957) mukaan termin frekvenssistä voidaan tehdä päätelmiä termin tärkeydestä kyseiselle dokumentille. Painofunktiona tf_{t_d} on yksinkertaisimmillaan termin t esiintymiskerrat dokumentissa d. Tyypillisesti raa'at frekvenssit normalisoidaan dokumentin pituuden suhteen, jotta vertailut eivät suosisi suuria dokumentteja.

Luonnollisten kielten yleiset sanat, kuten esimerkiksi suomen kielessä konjunktiot ja persoonapronominit, kuvaavat dokumenttien sisältöä joko hyvin vähän tai eivät ollenkaan. Fre-



Kuvio 1. Vektoriavaruusmalli (mukaillen Salton, Wong ja Yang 1975)

kvenssiin perustuva painotus antaisi näille sanoille todennäköisesti vektoriavaruusmallissa liian suuren painoarvon, kun painotetaan sisällöllisesti relevantteja termejä. Jones (1972) argumentoi, että useissa dokumenteissa yleiset termit ovat tarpeellisia, mutta niiden diskriminoiva voima on vähäinen. Jones ehdottaa painotusskeemaa, missä dokumenttijoukossa harvinaisia termejä painotettaisiin enemmän kuin yleisiä termejä. Tämä painofunktion nimeksi vakiintui myöhemmin IDF (inverse document frequency). IDF-painotuksessa termin arvokkuus on kääntäen verrannollinen termin dokumenttifrekvenssiin. IDF-funktion viitatuin muoto määritellään seuraavasti:

$$idf(t_i) = log\frac{N}{n_i}, \tag{2.2}$$

missä N on dokumenttien lukumäärä ja n_i on niiden dokumenttien lukumäärä missä, termi t_i esiintyy.

Dokumenttien kannalta kuvaavimmat termit ovat Saltonin ja Yangin (1988) mukaan mahdollisimman diskriminoivia muiden dokumenttien suhteen, toisin sanoen sellaiset termit, joilla on alhainen frekvenssi koko dokumenttijoukossa, mutta kuitenkin korkea frekvenssi pienessä dokumenttien osajoukossa. Yllä kuvattujen painofunktioiden tulo $tf \times idf$ korostaa näitä termejä (Salton ja Yang 1973). Saltonin ja Buckleyn (1988) mukaan tätä yhdistettyä painofunktiota $tf \times idf$ voidaan käyttää termin tärkeyden mittarina dokumenteille.

2.2 Aineiston esikäsittely

Dokumenttivektoreita käsitellään tyypillisesti termi-dokumenttimatriiseina, missä termit vastaavat matriisin rivejä ja dokumentit sarakkeita. Näitä termi-dokumenttimatriiseja tyypillisesti esikäsitellään ennen analysointia, sovelluskohteen määrittämällä tavalla. Esikäsittelyllä voidaan tavoitella parempaa representaatiota, vähäisempää kohinaa tai pienempää aineiston dimensionaalisuutta.

Luhnin (1960) mukaan dokumenttien sisältämien termien voidaan nähdä olevan representaation kannalta merkityksellisiä tai merkityksettömiä; yleisten sanojen voidaan turvallisesti ajatella olevan vähäarvoisia ja siten ne voidaan poistaa aineistosta. Suomen kielessä esimerkiksi sanaluokkien, kuten konjunktioiden ja pronominien yleisten sanojen voidaan todeta olevan sisällön kuvaavuuden kannalta merkityksettömiä, joten niiden poistaminen aineistosta ei todennäköisesti vaikuta dokumenttien vertailutarkkuuteen. Merkityksettömien sanojen (eng. stop words) poistaminen on yksi yleinen esikäsittelyvaihe. Yleisten sanojen poistamisella ei ole ensisijaisesti vaikutusta representaatioon, mutta se vähentää aineiston dimensioita ja siten nopeuttaa laskentaa.

Vektoriavaruusmallissa sanojen eri taivutusmuodot tulkitaan eri termeiksi, mikä aiheuttaa epätarkkuutta dokumenttien vertailussa. Stemmauksella tarkoitetaan tekniikoita, jotka pyrkivät muuntamaan taivutetut sanat niiden vartaloon, poistamalla sanasta sanaliitteet eli affiksit ja johtimet. Sanojen vartalo voi olla eri kuin sanan morfologinen perusmuoto eli lemma. Sanan muuntamista perusmuotoon kutsutaan lemmaukseksi. Suomenkielisissä aineistoissa stemmauksen ja lemmauksen merkitys on huomattavasti suurempi, kuin esimerkiksi englanninkielisissä aineistossa. Tämä johtuu siitä, että suomen kielen morfologia on huomattavasti

laajempi kuin englannin. Eri taivutusmuodot tulkitaan vektoriavaruusmallissa eri termeiksi, joten suomenkieliset aineistot tuottavat todennäköisesti pidempiä dokumenttivektoreita, kuin englanninkieliset aineistot.

Songin, Liun ja Yangin (2005) tuloksien mukaan yleisten sanojen poistamisen ja stemmauksen vaikutus luokittelutarkkuuteen on vähäinen. Dokumenttivektoreiden normalisoinnilla oli kuitenkin näkyvästi positiivinen vaikutus luokittelutarkkuuteen kaikissa testatuissa aineistoissa. Stemmausta heidän mukaansa voidaan suositella luokittelussa, koska stemmauksen avulla voidaan vähentää aineiston dimensioita huomattavasti.

2.3 Semanttinen representaatio

Vektoriavaruusmalli on täysin syntaksiin pohjautuva representaatio, se siis kadottaa alkuperäisistä dokumenteista sanajärjestyksen ja lauserakenteet. Dokumentit esitetään vektoriavaruusmallissa yksinkertaisesti termien joukkoina, tästä syystä malliin viitataan usein BoW-mallina (Bag-of-Words -mallina). BoW-malliin perustuvia representaatioita vaivaa erityisesti kaksi ongelmaa: synonymia ja homonymia. Synonymialla viitataan tässä tutkielmassa yleisesti siihen ilmiöön, että syntaktisesti erilaisilla termeillä voidaan viitata samaan informaatioon. Homonymialla taas viitataan tässä kontekstissa siihen, että syntaktisesti samalla termillä voi olla useita eri merkityksiä. Varsinkin kun dokumentit ovat lyhyitä, yllä mainitut ongelmat korostuvat, sillä satunnaisilla sanavalinnoilla on suurempi merkitys dokumenttien vertailussa.

Dokumenttien samankaltaisuuden vertailu perustuu BoW-mallissa termien esiintyvyyksien vertailuun, esimerkiksi dokumenttivektoreiden pistetulo kaavassa (2.1). Synonyymisten termien takia samankaltaisuuskerroin voi olla todellista samankaltaisuutta alhaisempi ja homonymian taas voi aiheuttaa samankaltaisuusmittauksessa vääriä positiivisia havaintoja. Sanavalinnat siis aiheuttavat dokumenttien vertailussa epävarmuutta. Tätä ongelmaa on pyritty ratkaisemaan etsimällä dokumenteille representaatiota, joka kykenisi mallintamaan dokumenttia konseptuaalisella tasolla, siis semanttista representaatioita.

Semanttisen representaation lähestymistapoja voidaan Cambrian ja Whiten (2014) mukaan luokitella karkeasti kahteen ryhmään: ontologiseen lähestymistapaan ja laskennalliseen lä-

hestymistapaan. Ontologisessa lähestymistavassa hyödynnetään ulkopuolisia tietokantoja tai tiedonlähteitä, kuten WordNet tai Wikipedia. Laskennallisessa lähestymistavassa taas hyödynnetään dokumenttien sisäistä semanttista rakennetta piirteiden löytämiseksi. Pääakselihajotelma (eng. singular value decomposition) on eräs laskennallinen menetelmä, jonka avulla voidaan löytää dokumenttiaineistosta konseptuaalisia piirteitä (Deerwester ym. 1990).

3 Pääakselihajotelma

Vektoriavaruusmallin mukaisen representaation voidaan olettaa sisältävän sanavalinnoista johtuvaa kohinaa. Pääakselihajotelmalla termi-dokumenttiaineisto voidaan muuntaa konseptiavaruuteen, jolloin termien ja dokumenttien väliset assosiatiiviset suhteet tulevat esiin. Termi-dokumenttiaineisto muunnetaan konseptiavaruuteen approksimoimalla aineistoa pääakselihajotelmalla löydettävillä konsepteilla. Tähän prosessiin viitataan tutkielmassa pääakselihajotelma-approksimaatiolla.

Pääakselihajotelma-approksimaatiossa alkuperäinen termi-dokumenttimatriisi esitetään pääakselihajotelman avulla kolmen erityisen matriisin tulona. Nämä kolme matriisia sisältävät alkuperäisen matriisin ominaisvektoreita ja ominaisarvoja. Matriiseissa nähdään alkuperäisen matriisin hajotelma lineaarisesti riippumattomiksi faktoreiksi. Osa näistä faktoreista on hyvin pieniä, eli niiden merkitys alkuperäiselle matriisille on vähäinen. Nämä pienet faktorit voidaan nollata ja siten muodostaa approksimaatio alkuperäisestä matriisista. Näihin faktoreihin viitataan usein myös konsepteina, sillä ne kuvaavat termien ja dokumenttien assosiatiivisia suhteita. Aineiston approksimointiin käytetään pientä osaa konsepteista, jotka selittävät suurta osaa aineiston datasta. Approksimoidussa aineistossa dokumentit representoidaan näiden konseptien avulla. Approksimoinnin vaikutuksena toisiinsa liittyvät dokumentit ja termit voivat kuvautua lähelle toisiaan konseptiavaruudessa, joka mahdollistaa paremman representaation. Termi-dokumenttimatriisin approksimointiin pääakselihajotelmalla viitataan myös nimellä LSI (latent semantic indexing). (Deerwester ym. 1990).

Olkoon $m \times n$ termi-dokumenttimatriisi X, jonka aste on r. On olemassa matriisin X pääakselihajotelma, joka on määritelty seuraavasti:

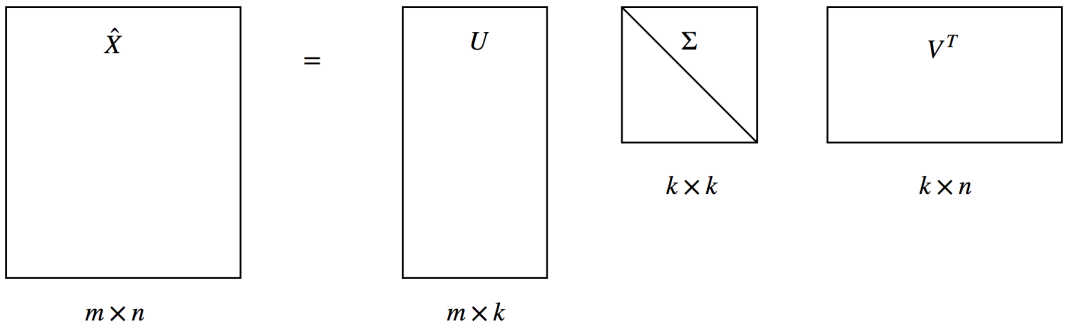
$$X = U\Sigma V^T, \tag{3.1}$$

missä U ja V ovat matriiseja, joiden sarakkeet ovat ortonormaaleja, joista U sisältää vasemmat ominaisvektorit ja V sisältää oikeat ominaisvektorit. Σ on diagonaalimatriisi, joka sisältää X:n ominaisarvot $\sigma_1, \sigma_2, \ldots \sigma_n$ siten, että $\sigma_i \geq \sigma_{i+1} \geq 0$. Koska X:n ominaisarvot ovat

järjestetty laskevaan järjestykseen, voidaan matriisia X approksimoida valitsemalla Σ:sta k suurinta ominaisarvoa ja nollata loput. Laskennan yksinkertaistamiseksi nollatut rivit ja sarakkeet voidaan poistaa Σ:sta, sekä myös nollattuja ominaisarvoja vastaavat U:n ja V:n sarakkeet. Näiden matriisien tulona saadaan approksimaatio $\hat{X}$:

$$X \approx \hat{X} = U_k \Sigma_k V_k^T, \tag{3.2}$$

missä Σ_k on $k \times k$ diagonaalimatriisi, U_k sisältää k saraketta ja V_k^T sisältää k riviä. Approksimoidun aineiston $\hat{X}$:n aste on $k \leq r$. Approksimaation tasoa voidaan säätää k-parametria muuntamalla, mitä kauempana k on r:stä, sitä vähemmän faktoreita käytetään approksimaation muodostamiseen ja sitä epätarkempi approksimaatio on. Kaavan 3.2 k:nen asteen pääakselihajotelmaa kutsutaan katkaistuksi pääakselihajotelmaksi (eng. truncated singular value decomposition).



Kuvio 2. Katkaistu pääakselihajotelma termi-dokumenttimatriisille X (mukaillen Deerwester ym. 1990)

Kuviossa 2 Matriisin X pääakselihajotelma-approksimaatio esitettynä graafisesti, missä:

U:n sarakkeet ovat ortonormaaleja $(U^T U = I)$
V:n sarakkeet ovat ortonormaaleja $(V^T V = I)$
Σ on diagonaalimatriisi, joka sisältää X:n ominaisarvoja
m on X:n rivien määrä
n on X:n sarakkeiden määrä
r on X:n aste $(r \leq min(m,n))$

k on valittu approksimaatiotaso $(k \leq r)$

Approksimoitua aineistoa $\hat{X}$ voidaan päivittää laskematta pääakselihajotelmaa uudestaan koko aineistolle projisoimalla uudet dokumenttivektorit valmiiseen aliavaruuteen (eng. folding-in). Tämä voi olla haluttavaa, sillä pääakselihajotelman laskeminen varsinkin suurille aineistoille voi olla hidasta. Uuden dokumenttivektorin projektio $\hat{d}$ aliavaruuteen lasketaan seuraavasti:

$$\hat{d} = d^T U_k \Sigma_k^{-1}, \tag{3.3}$$

missä d on $m \times 1$ dokumenttivektori ja U_k sisältää $\hat{X}$:n termivektorit (Berry, Dumais ja O'Brien 1995). Projektion jälkeen $\hat{d}$ voidaan liittää osaksi $\hat{X}$:a. Aineiston päivittäminen projisoiduilla dokumenttivektoreilla on huomattavasti laskennallisesti nopeampaa, kuin pääakselihajotelman laskeminen koko aineistolle uudestaan. Projektiomenetelmän käyttö voi kuitenkin heikentää representaatiota, koska konseptit eivät voi sovittautua uuteen dataan (Berry, Dumais ja O'Brien 1995).

Kaavan 3.3 projektiomenetelmää voidaan representaation päivittämisen lisäksi käyttää luokittelussa käytettävien testiaineistojen projisointiin samaan aliavaruuteen koulutusaineiston kanssa. Tämä on välttämätöntä, jotta koulutusaineistoon sovitettu luokittelualgoritmi toimisi testiaineiston kanssa. Koska projektiolla luotu representaatio voidaan olettaa heikommaksi kuin pääakselihajotelmalla approksimoitu representaatio, voidaan siten projisoidun testiaineiston representaatio olettaa epäoptimaaliseksi. Jotta testiaineiston representaatio ei olisi merkittävästi heikompi kuin koulutusaineiston representaatio, tulisi pääakselihajotelmalla olla runsaasti dataa, jotta tekstiaineistosta löytyvät termien ja dokumenttien väliset yhteydet selittyisivät mahdollisimman kattavasti koulutusaineiston konsepteilla. Pääakselihajotelmalla approksimoidun representaation päivittämiseen on kaavan 3.3 lisäksi robustimpi menetelmä, joka mukauttaa alkuperäistä representaatiota uuteen dataan säilyttämällä ortogonaalisuuden (ks. O'Brien 1994).

Aineiston approksimointi katkaistulla pääakselihajotelmalla vähentää aineiston astetta $(k \leq r)$, approksimoidun aineiston dimensiot kuitenkin pysyvät ennallaan. Mikäli piirteiden mää-

rää halutaan vähentää approksimoidussa aineistossa, voidaan aineiston alkuperäiset dokumenttivektorit projisoida haluttuun mittaan kaavaa 3.3 hyödyntäen. Aineisto voidaan täten approksimoida $k \times n$ matriisilla $\hat{X}$.

Deerwester ym. (1990) mukaan pääakselihajotelmalla löydettävä konseptipohjainen representaatio on eräs ratkaisu synonyymisten termien ongelmaan. Useiden merkitysten ongelman suhteen menetelmä ei heidän mukaansa kuitenkaan ole yhtä onnistunut, sillä termien merkitykset eivät voi teknisesti kuvautua kuin yhteen pisteeseen konseptiavaruudessa. Heidän mukaansa termit jolla on useita eri merkityksiä, esitetään konseptiavaruudessa näiden merkitysten keskiarvona. Homonyymiset termit voivat heidän mukaansa aiheuttaa representaatioon vääristymiä, mikäli nämä oikeat merkitykset eivät sijaitse lähellä keskiarvoista merkitystä.

4 Katkaistu pääakselihajotelma luokittelussa

Tekstin luokittelussa haasteita aiheuttavat aineiston suuri dimensionalisuus, harva data sekä useat epärelevantit piirteet. Näistä syistä aineiston esikäsittely on tyypillistä ennen luokittelualgoritmin sovittamista. Luvussa 2.2 esittelemien esikäsittelytekniikoiden lisäksi aineistoa voidaan rajata (eng. feature selection) ennen luokittelua. Aineiston rajauksen tarkoituksena on valita aineiston piirteiden joukosta sellaiset piirteet, joiden avulla saavutetaan käytössä olevalla luokittelualgoritmilla paras luokittelutarkkuus. Hyvällä aineiston rajauksen menetelmällä voidaan parantaa luokittelutehokkuutta sekä tarkkuutta. Yangin ja Pedersenin mukaan (1997) tekstin luokittelussa esimerkiksi dokumenttifrekvenssi on yllättävän tehokas aineiston rajausmenetelmä. Tutkimuksessaan he osoittivat, että dokumenttifrekvenssi korreloi vahvasti muiden tehokkaiden rajausmenetelmien kanssa. Tämä heidän mukaansa viittaa siihen, että yleiset termit ovat tärkeitä tekstin luokittelussa.

Pääakselihajotelmalla approksimointi on myös aineistoa rajaava menetelmä, sillä pääakselihajotelmalla voidaan uudelleenparametrisoida aineisto valittua k:ta konseptia käyttäen. Nämä valitut konseptit ovat termien ja dokumenttien välisiä assosiatiivisia yhteyksiä, joiden merkittävyys on aineistossa suurin. Hyvän approksimaation luomisessa k-parametrin valinta on kriittistä. Approksimaation taso tulisi olla riittävän korkea, että aineiston representaatio ei heikkene, mutta myös riittävän alhainen, jotta kohina vaimenee. Deerwester ym. (1990) mukaan konseptien lisääminen nostaa approksimoidun representaation tasoa tiettyyn pisteeseen asti, kunnes konseptien lisääminen laskee representaation tasoa. Tätä ilmiötä voidaan heidän mukaansa tulkita ylisovittumisena, ylimääräisten konseptien lisääminen mallintaisi aineiston rakenteen sijaan aineistossa esiintyvää kohinaa.

Luonnollinen kieli on luonteeltaan hyvin vaihtelevaa, joten sovelluksissa on löydettävä aineistokohtaisesti approksimaation taso, joka maksimoi luokittelutarkkuuden. Tämä voi todistautua ongelmalliseksi, sillä tiedossa ei ole automaattisia keinoja määrittää optimaalista konseptien määrää, joilla aineisto approksimoidaan. Tämä tarkoittaa sitä, että optimaalinen approksimaatiotaso täytyy löytää vertaamalla eri tason approksimaatioita, joka voi aineiston koosta riippuen olla hyvin työlästä. Aineiston vaihtelun vuoksi suositusta konseptien määräksi on vaikea antaa. Deerwester ym. (1990) kuitenkin arvelevat 50 – 150 konseptin riittävän

hyvään representaatioon. Mengin, Lin ja Yun (2011) tutkimuksessa korkein luokittelutarkkuus saavutettiin aineistosta ja rajausmenetelmästä riippuen 100 – 150 konseptilla.

Luonnollisessa kielessä erilaisilta dokumenttityypeiltä voidaan odottaa eri kielen rakennetta ja sanavalintoja. Tämä ilmiö näkyy luonnollisen kielen prosessoinnissa eri representaatiomenetelmien vaikutusten vaihteluna eri aineistojen välillä (Song, Liu ja Yang 2005). Aineistojen vaihtelulla näyttää myös olevan suuri merkitys approksimoidun representaation tasoon. Deerwester ym. (1990) mukaan suuridimensioinen ja runsas termi-dokumenttiaineisto näyttää olevan tarpeellinen hyvään konseptipohjaiseen representaatioon. He käyttivät tutkimuksessaan MED ja CISI aineistoja. MED aineisto sisältää 1033 lääketieteen abstraktia ja 5823 termiä ja CISI aineisto sisältää 1460 informaatiotieteen abstraktia ja 5135 termiä. Yleiset termit olivat suodatettu aineistoista. Aineistojen approksimointiin käytettiin sataa konseptia. Approksimoinnista oli heidän mukaansa hyötyä MED aineistossa, CISI aineistossa approksimoinnin vaikutukset olivat vähäisiä.

Pääakselihajotelman hyödyntäminen näyttää vaativan runsaasti aineistoa, jotta menetelmällä voidaan löytää merkityksellisiä konsepteja aineistosta. Konsepteja voidaankin ajatella tilastollisesti semanttisina piirteinä, sillä ne kuvaavat joukkoa termejä ja dokumentteja jotka esiintyvät usein lähellä toisiaan. Dataa täytyy olla tarpeeksi, jotta aineistossa oleva rakenne tulee esiin. Jos aineistoa on riittävästi pääakselihajotelman hyödyntämiseen, voidaan approksimoidulla representaatiolla saada parempia luokittelutuloksia, kuin perinteisillä BoW-representaatiolla. Zhang, Yoshida ja Tang (2011) vertasivat pääakselihajotelmalla approksimoidun representaation vaikutusta luokittelutarkkuuteen. Tutkimuksessa käytettiin kiinankielistä TanCorp1.0 aineistoa ja englanninkielistä Reuters-21578 aineistoa. TanCorp1.0 sisältää dokumentteja kiinankielisistä akateemisista julkaisuista, aineistosta valittiin yhteensä 1200 dokumenttia neljästä kategoriasta. Englanninkielinen Reuters-21578 aineisto sisältää uutisia brittiläisestä Reuters julkaisusta, josta valittiin yhteensä 2042 dokumenttia neljästä kategoriasta. Approksimoidulla aineistolla saavutettiin merkittävästi ($P \leq 0.01$) tarkempi luokittelutarkkuus molemmissa aineistossa, verrattuna $tf \times idf$ representaatioon.

Mengin, Lin ja Yun (2011) mukaan tekstin luokittelussa suuri termien määrä todennäköisesti aiheuttaa luokittelualgoritmin ylisovittumisen. Luokittelutarkkuuden parantamiseksi he ehdottavat aineiston rajaamista ennen pääakselihajotelma-aproksimaation suorittamista. Tut-

kimuksessaan Meng, Lin ja Yu (2011) vertaavat eri menetelmiä valita luokitteluaineistosta tärkeimpiä piirteitä ennen pääakselihajotelman soveltamista. Tätä menetelmää he kutsuvat kaksivaiheiseksi aineiston rajaukseksi. Kaksivaiheisessa aineiston rajauksessa siis luokitteluaineistosta valitaan käytetyllä rajausmenetelmällä diskriminoivimmat piirteet, jonka jälkeen aineisto muunnetaan konseptuaaliseen avaruuteen pääakselihajotelmalla approksimoimalla. Heidän tutkimuksessaan esitetään lupaavia tuloksia kaksivaiheisen aineiston rajaamisen puolesta. Menetelmän mahdollinen luokittelutarkkuutta edistävä vaikutus voi viitata siihen, että aineiston rajaaminen edistää diskriminoivien konseptien muodostusta. Menetelmän ja sen vaikutusta pääakselihajotelmalla muodostettaviin konsepteihin olisi kuitenkin vielä syytä tutkia tarkemmin.

Pääakselihajotelmalla approksimoimalla voidaan vähentää aineistossa olevaa sanavalinnoista johtuvaa kohinaa. Menetelmä mahdollistaa vertailut dokumenteille, jotka eivät sisällä samoja termejä. Pääakselihajotelmalla approksimointi myös rajaa aineistoa valittuun määrään konsepteja. Pääakselihajotelma-approksimaatio ei kuitenkaan ole luokittelun kannalta optimaalinen menetelmä, sillä pääakselihajotelma ei ota kantaa dokumenttien luokkiin approksimaation yhteydessä (Sun ym. 2004; Cai, He ja Han 2005). Menetelmässä suurimmat konseptit ovat termien ja dokumenttien välisiä assosiatiivisia yhteyksiä, jotka selittävät alkuperäisen aineiston dataa eniten. Pääakselihajotelma-approksimaatio ei siis valitse mahdollisimman diskriminoivia piirteitä. Tätä ongelmaa on pyritty ratkaisemaan menetelmällä, joka valitsee pääakselihajotelmasta diskriminoivimmat kantavektorit, joita käytetään dokumenttivektoreiden projisointiin (ks. Sun ym. 2004).

Deerwester ym. (1990) ehdottivat termi-dokumenttiaineiston approksimointia katkaistulla pääakselihajotelmalla alun perin tiedonhaun tarpeisiin. Menetelmällä oli tarkoitus löytää parempi representaatio tekstiaineistoille, sillä vektoriavaruusmallia vaivaavat kaksi perustavanlaatuista ongelmaa, synonymia ja homonymia. Zhangin, Yoshidan ja Tangin (2011), ja Mengin, Lin ja Yun (2011) tutkimustulokset antavat näyttöä siitä, että katkaistu pääakselihajotelma soveltuu tekstin representointiin myös luokittelussa, mutta kuten useimmissa tekstin representaation menetelmissä, käytettävä aineisto vaikuttaa suuresti tämänkin menetelmän tehokkuuteen. Pientä aineistoa approksimoitaessa on vaarana, että satunnaisuuden vaikutus konsepteihin on suuri. Tästä syystä hyvä representaatio näyttää vaativan runsaasti dataa, jotta

aineistosta löytyvä rakenne tulee esiin.

Deerwester ym. (1990) mukaan pääakselihajotelmalla approksimointi mahdollistaa synonyymisten sanojen paremman representaation, kuitenkaan homonyymisten termien ongelmaa menetelmä ei ratkaise täysin. Tämä johtuu siitä, että eri merkitykset voivat kuvautua vain yhteen pisteeseen avaruudella. Tarkka homonyymisten sanojen representointi on haastavaa, sillä, se vaatisi kontekstisidonnaisten termien merkitysten tunnistamisen ja kategorisoinnin. Homonyymisten termien epäoptimaalinen representaatio lienee kuitenkin hyväksyttävä kompromissi, sillä synonyymiset termit voidaan representoida pääakselihajotelma avulla paremmin. Homonymian vaikutusta pääakselihajotelmalla approksimointuun representaatioon olisi kuitenkin syytä tutkia tarkemmin.

5 Yhteenveto

Tekstiaineistojen luokittelun tekee haasteelliseksi dimensionaalisuus, harva data ja useat epärelevantit piirteet. Tekstin syntaktisen representaation voidaan olettaa sisältävän sanavalinnoista johtuvaa kohinaa, mikä tekee dokumenttien vertailusta haastavaa. Näitä haasteita on pyritty ratkaisemaan kehittämällä erilaisia esikäsittelymenetelmiä, kuten yleisten sanojen suodattaminen ja stemmaus. Sanavalinnasta johtuvaa kohinaa ei kuitenkaan voida täysin esikäsittelymenetelmillä ratkaista. Termien eri taivutusmuodot voidaan muuntaa niiden yhteiseen vartaloon stemmauksella, mutta synonyymiset ja homonyymiset termit voivat silti aiheuttaa dokumenttien vertailuissa vääristymiä.

Tekstiaineiston kohinaa voidaan vaimentaa uudelleenparametrisoimalla aineisto katkaistulla pääakselihajotelmalla approksimoimalla. Approksimaatiossa aineisto representoidaan pääakselihajotelmalla löydettävillä konsepteilla, jotka kuvaavat termien ja dokumenttien assosiatiivisia yhteyksiä. Uudelleenparametrisointi mahdollistaa samankaltaisuuksien löytämisen dokumenttien välille, jotka eivät sisällä yhtään samoja termejä.

Aineistoa approksimoimalla voidaan pienentää aineiston dimensiota ja vaimentaa aineiston kohinaa, samalla edistäen luokittelutarkkuutta. Pääakselihajotelma ei ole kuitenkaan menetelmänä ongelmaton. Hyvän representaation muodostus näyttää vaativan runsaasti dataa. Representaation tasoon vaikuttaa myös olennaisesti approksimointiin käytettävien konseptien määrä, joka voi olla työlästä optimoida. Pääakselihajotelma ei myöskään valitse approksimointiin diskriminoivimpia konsepteja, joka tekee menetelmästä epäoptimaalisen luokittelun kannalta. Yhteenvetona voidaankin todeta, että pääakselihajotelmaa ei pidä käyttää kaiken kattavana menetelmänä, vaan yhtenä mahdollisena osana tekstin representaation kokonaisuutta.

Luonnollisen kielen mallintaminen syntaktisesti luo väistämättä representaatioon vääristymiä, sillä luonnolliseen kieleen liittyy vahvasti tulkinnanvaraisuutta. Jatkotutkimuksen kannalta olisi siksi kiinnostavaa nähdä systemaattisia vertailuja eri konseptipohjaisille representaatiomenetelmille. Tällaisia menetelmiä ovat esimerkiksi: LDA (latent Dirichlet allocation) ja läheisesti pääakselihajotelmaan liittyvät PLSA (probabilistic latent semantic analysis) ja

SLSI (supervised latent semantic indexing) (ks. Blei, Ng ja Jordan 2003; Hofmann 2001; Sun ym. 2004).

Katkaistun pääakselihajotelman käyttö representaatiomenetelmänä vaikuttaa lupaavalta, mutta vaatii vielä tutkimusta. Menetelmään liittyviä avoimia kysymyksiä ovat vielä esimerkiksi: miten eri esikäsittelymenetelmät ja painofunktiot vaikuttavat konseptien muodostukseen, millainen vaikutus aineiston rajauksella on approksimointiin ja millaisia vaikutuksia kielellä, kuten suomen kielellä, on representaatiomenetelmään.

Lähteet

Berry, M., S. Dumais ja G. O'Brien. 1995. "Using Linear Algebra for Intelligent Information Retrieval". *SIAM Review* 37 (4): 573–595. doi:10.1137/1037127. eprint: https://doi.org/10.1137/1037127. https://doi.org/10.1137/1037127.

Blei, David M, Andrew Y Ng ja Michael I Jordan. 2003. "Latent dirichlet allocation". *Journal of machine Learning research* 3 (Jan): 993–1022.

Cai, D., X. He ja J. Han. 2005. "Document clustering using locality preserving indexing". *IEEE Transactions on Knowledge and Data Engineering* 17, numero 12 (joulukuu): 1624–1637. ISSN: 1041-4347. doi:10.1109/TKDE.2005.198.

Cambria, E., ja B. White. 2014. "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]". *IEEE Computational Intelligence Magazine* 9, numero 2 (toukokuu): 48–57. ISSN: 1556-603X. doi:10.1109/MCI.2014.2307227.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer ja Richard Harshman. 1990. "Indexing by latent semantic analysis". *Journal of the American Society for Information Science* 41 (6): 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9.

Hofmann, Thomas. 2001. "Unsupervised Learning by Probabilistic Latent Semantic Analysis". *Machine Learning* 42, numero 1 (tammikuu): 177–196. ISSN: 1573-0565. doi:10.1023/A:1007617005950. https://doi.org/10.1023/A:1007617005950.

Jones, Karen Spärck. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval". *Journal of Documentation* 28 (1): 11–21. doi:10.1108/eb026526. eprint: https://doi.org/10.1108/eb026526. https://doi.org/10.1108/eb026526.

Luhn, H. P. 1957. “A Statistical Approach to Mechanized Encoding and Searching of Literary Information”. *IBM Journal of Research and Development* 1, numero 4 (lokakuu): 309–317. ISSN: 0018-8646. doi:10.1147/rd.14.0309.

———. 1960. “Key word-in-context index for technical literature (kwic index)”. *American Documentation* 11 (4): 288–295. doi:10.1002/asi.5090110403. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.5090110403. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090110403.

Meng, Jiana, Hongfei Lin ja Yuhai Yu. 2011. “A two-stage feature selection method for text categorization”. Computers & Mathematics in Natural Computation and Knowledge Discovery, *Computers & Mathematics with Applications* 62 (7): 2793–2800. ISSN: 0898-1221. doi:https://doi.org/10.1016/j.camwa.2011.07.045. http://www.sciencedirect.com/science/article/pii/S089812211100616X.

O’Brien, Gavin W. 1994. “Information management tools for updating an SVD-encoded indexing scheme”. Tutkielma, University of Tennessee, Knoxville.

Salton, G., A. Wong ja C. S. Yang. 1975. “A Vector Space Model for Automatic Indexing”. *Commun. ACM* (New York, NY, USA) 18, numero 11 (marraskuu): 613–620. ISSN: 0001-0782. doi:10.1145/361219.361220. http://doi.acm.org/10.1145/361219.361220.

Salton, G., ja C.S. Yang. 1973. “On the specification of term values in automatic indexing”. *Journal of Documentation* 29 (4): 351–372. doi:10.1108/eb026562. eprint: https://doi.org/10.1108/eb026562. https://doi.org/10.1108/eb026562.

Salton, Gerard, ja Christopher Buckley. 1988. “Term-weighting Approaches in Automatic Text Retrieval”. *Inf. Process. Manage.* (Tarrytown, NY, USA) 24, numero 5 (elokuu): 513–523. ISSN: 0306-4573. doi:10.1016/0306-4573(88)90021-0. http://dx.doi.org/10.1016/0306-4573(88)90021-0.

Song, Fengxi, Shuhai Liu ja Jingyu Yang. 2005. “A comparative study on text representation schemes in text categorization”. *Pattern Analysis and Applications* 8, numero 1 (syyskuu): 199–209. ISSN: 1433-755X. doi:10.1007/s10044-005-0256-3. https://doi.org/10.1007/s10044-005-0256-3.

Sun, Jian-Tao, Zheng Chen, Hua-Jun Zeng, Yu-Chang Lu, Chun-Yi Shi ja Wei-Ying Ma. 2004. “Supervised latent semantic indexing for document categorization”. Teoksessa *Fourth IEEE International Conference on Data Mining (ICDM’04),* 535–538. IEEE.

Yang, Yiming, ja Jan O. Pedersen. 1997. “A Comparative Study on Feature Selection in Text Categorization”. Teoksessa *Proceedings of the Fourteenth International Conference on Machine Learning,* 412–420. ICML ’97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1-55860-486-3. http://dl.acm.org/citation.cfm?id=645526.657137.

Zhang, Wen, Taketoshi Yoshida ja Xijin Tang. 2011. “A comparative study of TF*IDF, LSI and multi-words for text classification”. *Expert Systems with Applications* 38 (3): 2758–2765. ISSN: 0957-4174. doi:https://doi.org/10.1016/j.eswa.2010.08.066. http://www.sciencedirect.com/science/article/pii/S0957417410008626.