Etäisyysmittojen vertailu elektronisessa täsmäjakeluympäristössä

Matematiikan erikoistyöt Mat-1.139

Eerika Savia, 39482a, Tf
23.3.1998

Sisältö

1 Johdanto **2**

2 Metadatan rakenne **2**
- 2.1 Tausta 2
- 2.2 Uudenlainen metadata 3

3 Käyttäjäprofiilin rakenne **4**

4 Etäisyysmitan käsite **5**

5 Profiilin ja dokumentin välinen mitta **6**
- 5.1 Painokertoimet 8
- 5.2 Symmetrinen mitta 9
  - 5.2.1 Esimerkki 9
  - 5.2.2 Metriikan ehtojen toteutumisesta symmetriselle mitalle 13
- 5.3 Symmetrisyys 13

6 Karkea epäsymmetrinen etäisyysmitta **14**
- 6.1 Esimerkki 15
- 6.2 Metriikan ehtojen toteutumisesta karkealle mitalle 18

7 Tarkennusmitta karkealle mitalle **18**
- 7.1 Esimerkki 19
- 7.2 Metriikan ehtojen toteutumisesta tarkennusmitalle 19

8 Mittojen yhdistäminen **20**
- 8.1 Karkean mitan ja tarkennusmitan välinen yhteys 20
- 8.2 Yhdistetty mitta 21

9 Yhteenveto **24**

A Liite : Tuloksia **26**
- A.1 Profiili user1 26
- A.2 Toinen käyttäjäprofiili 28

1 Johdanto

Tämä työ on syntynyt osana TKK:n TAI-tutkimuslaitoksen projektia nimeltä SmartPush. Elektronisessa täsmäjakeluympäristössä käyttäjille pyritään tarjoamaan räätälöityä informaatiota digitaalisesta materiaalitulvasta, jonka hallitseminen ilman suodattamista olisi käyttäjille ylivoimaista tai vähintäänkin liian työlästä.

Suodattimina toimivat nk. *käyttäjäprofiilit*, joihin on tallennettu tietoa siitä, millaiset dokumentit kyseistä käyttäjää kiinnostavat [4]. *Metadata* kuvaa dokumentin olennaisia piirteitä, kuten sisältöä, julkaisijaa, ajankohtaisuutta ja tyyppiä (esim. kuva, ääni, teksti tai multimedia). Metadatan sisältö ja rakenne ovat myös SmartPush-projektin tutkimuskohteena. Metadatan avulla voidaan automaattisesti päätellä, millaista dokumenttivalikoimaa kullekin käyttäjälle kannattaisi tarjota.

Jotta profiili voisi suodattaa sitä vastaavia dokumentteja käyttäjälle, täytyy luonnollisesti määrätä etäisyysmitta, joka mittaa profiilin ja dokumentin välistä vastaavuutta. Tässä työssä vertaillaan erilaisia mittoja ja kehitetään SmartPush-ympäristöön soveltuva mitta.

2 Metadatan rakenne

2.1 Tausta

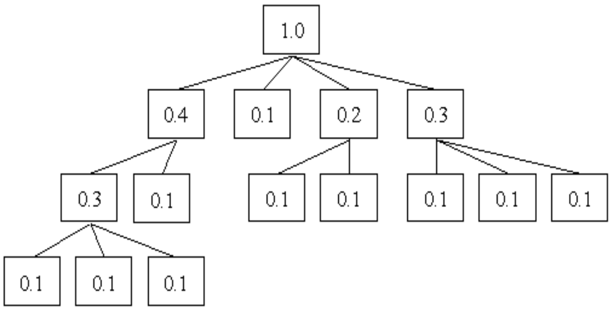
Tekstin rakenteeseen ja sanojen esiintymisfrekvensseihin perustuen on aiemmin toteutettu useita vastaavia täsmäjakeluympäristöjä [2]. Automaattisella kielentunnistuksella on pyritty poimimaan olennaiset avainsanat eli termit raakatekstistä ja dokumentti on siis luokiteltu pelkästään sisältämiensä sanojen perusteella. Dokumentin esitykseksi saadaan silloin vektori, jossa kullakin termillä on oma esiintymistiheydestä riippuva painonsa. Vektoriesityksessä on kuitenkin paljon redundanttia tietoa, koska termit eivät ole ortogonaalisia, vaan niiden esiintymistiheydet riippuvat toisistaan. Etäisyysmitat tyypillisesti perustuvat komponenttien ortogonaalisuuteen, joten redundantista tiedosta on muutakin haittaa kuin tarpeettoman suuri dokumenttiavaruuden dimensio.

Latentti semanttinen indeksointi [3] käyttää dokumenttidataan singulaariarvohajotelmaa dimension pienentämiseksi. Jatkuvalle dokumenttivirralle jouduttaisiin kuitenkin tekemään oletus, että tuleva virta vastaa jotain tiettyä otosta, jolle singulaariarvohajotelma on jo laskettu.

2.2 Uudenlainen metadata

Tekstissä esiintyviin avainsanoihin pohjautuva kuvaustapa soveltuu ainoastaan tekstidokumenteille, mutta SmartPush-projektissa kehitetään menetelmää, joka sopii myös kuville, äänelle, videodatalle ja multimediadokumenteille. Metadatan pitäisi antaa kattava kuvaus kaikenlaisille dokumenteille, jotta avulla voitaisiin löytää käyttäjäprofiilia vastaavaa materiaalia rajoittumatta tekstiin. Metadatassa voi olla useita komponentteja, joita kutsutaan *dimensioiksi.* Ne voivat olla rakenteeltaan erilaisia : diskreettejä, jatkuvia tai hierarkisia. Tässä työssä keskitytään hierarkisen (sisältö)dimension etäisyysmittaan.

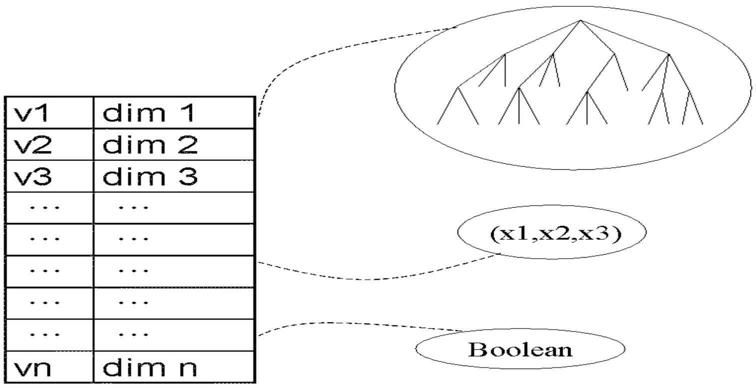
Dokumentin sisältödimensio voidaan vektorin sijasta esittää hierarkisena rakenteena, jossa puun lehdet vastaavat mallin tarkimpia käsitteitä. Hierarkinen rakenne luodaan olemassa olevien luokittelurakenteiden pohjalta. Aihesisältö esitetään puulla, jossa on painojakauma lehtitasolla (painojen summa on yksi), eli dokumentti kuuluu sumeasti eri luokkiin. Puun ylemmät solmut saavat painonsa suoraan alipuidensa painojen summana ja joka tasolla summa on siis yksi. Puun ylemmillä tasoilla päästään lähemmäs ortogonaalisuutta, kun samaan aiheeseen liittyvät käsitteet on niputettu samaan alipuuhun.



Kuva 1: Esimerkki metadatan sisältödimensiosta.

3 Käyttäjäprofiilin rakenne

Käyttäjäprofiilin on tarkoitus olla kompakti yhteenveto piirteistä, jotka määräävät dokumentin kiinnostavuuden kyseisen käyttäjän kohdalla. Käytännössä piirteet ovat niitä dimensioita, joita dokumenttien metadatassa esiintyy, painotettuna sen mukaan, mikä käyttäjää kiinnostaa. Käyttäjäprofiilissa on kullekin dimensiolle painokertoimet, jotka kuvaavat eri dimensioiden välistä tärkeysjärjestystä. Painot v_j ovat profiilikohtaiset ja $\sum_j v_j = 1$.



Kuva 2: Käyttäjäprofiilin rakenne.

Kuvassa 2 on kaavio käyttäjäprofiilista. Dimensioita dim 1 ... dim n vastaavat painot $v_1 \dots v_n$ ja jokaisen dimension yhteydessä on profiilin esitys, joka voi eri dimensioissa olla erilainen, esimerkiksi puurakenne, reaaliarvoinen vektori tai totuusarvo. (Käyttäjällä voisi myös olla useampi profiili eri käyttötarkoituksiin.) Käyttäjäprofiilin ja dokumentin välinen etäisyys muodostuu dimensiokohtaisista etäisyyksistä E_j, jotka on kaikki normeerattu välille $[0, 1]$.

$$D(u, d) = \sum_{j=1}^{n} v_j E_j(u, d) \quad , \tag{1}$$

missä u = käyttäjäprofiili ja d = dokumentin metadataesitys.

Käyttäjäprofiilissa sisältödimension esitys on vastaava kuin dokumenttienkin: puussa on painojakauma, joka kuvaa kiinnostuksen jakautumista eri aihepiireihin. (Painojen summa on yksi.) Puurakenteesta on se etu, että jo korkealla tasolla voidaan todeta, kannattaako käyttäjäprofiilia ja dokumenttia verrata tarkemmin, sillä puuta alaspäin mentäessä erotus voi vain kasvaa.

Tästä eteenpäin profiilin hierarkinen sisältödimensio samaistetaan käyttäjäprofiilin kanssa ja käsitellään siihen liittyvää etäisyysmittaa $E_j = E$.

4 Etäisyysmitan käsite

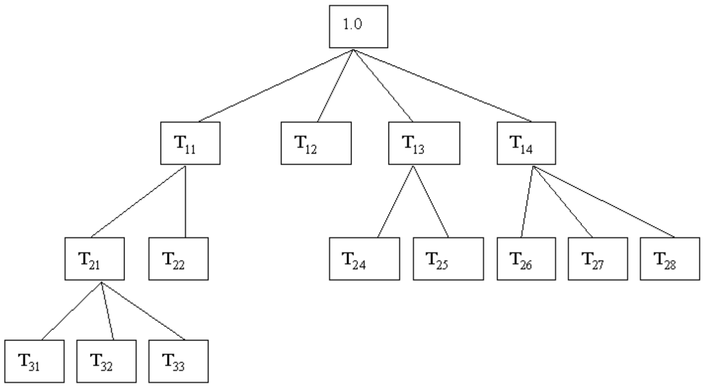
Määritelmä Olkoon S joukko. Kuvausta $d : S \times S \to \mathbf{R}$ sanotaan *metriikaksi* eli *etäisyydeksi* S:ssä, jos seuraavat ehdot ovat voimassa :

$$
\begin{array}{lll}
(i) & d(x,y) \geq 0 \quad \text{ja} \quad d(x,y) = 0 \iff x = y & \\
(ii) & d(x,y) = d(y,x) & \text{(symmetrisyys)} \\
(iii) & d(x,z) \leq d(x,y) + d(y,x) & \text{(kolmioepäyhtälö)}
\end{array}
$$

Kun mitataan käyttäjäprofiilin ja dokumentin välistä vastaavuutta, ei ole tarkoituksenmukaista vaatia, että mitta olisi aito metriikka. Esimerkiksi ehtoa (iii) ei tarvita ollenkaan, koska etäisyyksiä vain verrataan toisiinsa eikä missään tilanteissa lasketa yhteen. Kaikki verrattavat mitat ovat ei-negatiivisia, mutta ehdot (i) ja (ii) eivät muuten ole välttämättä voimassa. Etäisyysmitalla tarkoitetaan siis tässä työssä profiilin ja dokumentin välistä mittaa, joka ei välttämättä toteuta metriikan ehtoja. Eri ehtojen toteutumista tarkastellaan mittojen käsittelyn yhteydessä. Käyttäjäprofiilin ja dokumentin metadatan vertaamiseen tarvitaan mitta, jonka ensisijainen tarkoitus on muodostaa dokumenteille järjestys. Jotta dimensiokohtaiset etäisyydet E_j voidaan yhdistää kaavan (1) mukaisesti, tarvitaan myös järkevät lukuarvot etäisyyksille E_j.

5 Profiilin ja dokumentin välinen mitta

Merkitään puussa T tasolla m (juuresta laskien) ja omalla tasollaan indeksillä k sijaitsevaa solmua T_{mk}. Samaa notaatiota voidaan siis käyttää profiileille ja metadataesityksille. Voidaan siirtyä käyttämään solmun nimeä kyseisen solmun painon symbolina.



Kuva 3: Solmujen indeksointi.

Huomattavaa on, että vaikka jokin puun haara päättyisikin aikaisemmin, sen painoja ei unohdeta, vaan kaikki haarat voidaan ajatella jatkettavan puun alimmalle tasolle asti. Kuvan 3 puussa tämä tarkoittaa, että $T_{12} = T_{23} = T_{35}$, $T_{22} = T_{34}$, $T_{24} = T_{36}$, $T_{25} = T_{37}$, $T_{26} = T_{38}$, $T_{27} = T_{39}$ ja $T_{28} = T_{3,10}$.

Esimerkki
Kuvan 1 (sivu 3) painovektorit tasoittain ovat:

$$u_1 = (\ 0.4,\ 0.1,\ 0.2,\ 0.3\ )\ ,$$

$$u_2 = (\ 0.3,\ 0.1,\ 0.1,\ 0.1,\ \ 0.1,\ 0.1,\ 0.1,\ 0.1\ )\ ,$$

$$u_3 = (\ 0.1,\ 0.1,\ 0.1,\ 0.1,\ \ 0.1,\ 0.1,\ 0.1,\ 0.1,\ \ 0.1,\ 0.1\ ).$$

Käyttäjäprofiilin u ja dokumentin d välinen etäisyys $E(u, d)$ voidaan laskea esimerkiksi painotettuna summana eri tasoilla lasketuista etäisyyksistä :

$$E(u, d) = \sum_{m=1}^{N} w_m e_m(u_m, d_m) \quad , \tag{2}$$

missä m käy läpi kaikki puun tasot ja etäisyysmitta $e_m(\cdot, \cdot)$ voidaan valita vapaasti.

Lemma. Jos tasoittain laskettava mitta $e_m(u_m, d_m)$ on metriikka, niin kaavan (2) mitta $E(u, d)$ on myös metriikka, kun $w_N > 0$.
(missä N on puun lehtitaso).

Tod. Metriikan ehdot (ii) ja (iii) toteutuvat triviaalisti.
Olet. $e_m(u_m, d_m)$ on metriikka ja $w_m \geq 0 \quad \forall m$ ja $\sum_m w_m = 1$.
Ehto (i): Selvästi $E(u, d) \geq 0 \quad \forall u, d$.
Toinen suunta on selvä : $u = d \implies E(u, d) = 0$.
Oletetaan $E(u, d) = \sum_m w_m e_m(u_m, d_m) = 0$. Siitä seuraa, että on oltava $\forall m$ joko $w_m = 0$ tai $e_m(u_m, d_m) = 0$. Jos vaaditaan, että $w_m > 0 \quad \forall m$, niin $e_m(u_m, d_m) = 0 \implies u_m = d_m \forall m \implies u = d$.
Tässä sovelluksessa lehtitason ($m = N$) painot määräävät ylempien tasojen painot yksikäsitteisesti, joten $u_N = d_N \implies u_m = d_m \forall m$ ja riittävä ehto painoille on $w_N > 0$. $\square$

5.1 Painokertoimet

Painoilla w_m painotetaan eri tasoilla saatua etäisyyttä (kaavassa 2). Painojen summa normeerataan, jotta päästään vertaamaan eri painotusten tuottamia mittoja toisiinsa.

Painot voidaan valita esimerkiksi tasajakauman mukaisesti, jolloin kaikki tasot ovat samanarvoisia:

$$w_m = \frac{1}{N} \qquad m = 1, 2, \ldots, N. \tag{3}$$

Toisaalta painotusta voidaan kasvattaa puussa alaspäin mentäessä, jolloin tarkemmin samaan aiheeseen osuvat erottuvat selvemmin niistä, jotka osuvat vain ylemmillä tasoilla. Luonnollisesti jo tasajakauma painottaa hieman lehtitasoa, koska vaikutus kumuloituu joka tapauksessa puussa ylöspäin. Eräs esimerkki alaspäin kasvavasta painotuksesta on

$$\begin{cases} w_m = \left(\frac{1}{2}\right)^{N-m+1} & m = 2, 3, \ldots, N \\ w_1 = \left(\frac{1}{2}\right)^{N-1} & . \end{cases} \tag{4}$$

Painojen summaksi saadaan

$$\sum_m w_m = \sum_{m=1}^{N-1} \left(\frac{1}{2}\right)^m + \left(\frac{1}{2}\right)^{N-1} = \frac{1 - \left(\frac{1}{2}\right)^N}{1 - \frac{1}{2}} - 1 + \left(\frac{1}{2}\right)^{N-1} = 1 \; .$$

Vertailun vuoksi testidokumenteille on laskettu etäisyydet myös päinvastaisella painotuksella, jossa ylemmät tasot saavat isomman painon ja lehtitasolla on vähemmän painoarvoa. (Liite A.1)

$$\begin{cases} w_m = \left(\frac{1}{2}\right)^{m} & m = 1, 2, \ldots, N-1 \\ w_N = \left(\frac{1}{2}\right)^{N-1} & . \end{cases} \tag{5}$$

On oletettavaa, että tällainen painotus ei tuota hyvää dokumenttivalikoimaa käyttäjälle, koska se ei arvosta tarkkaa osumaa, toisin kuin lehtitason merkityksen korostaminen.

5.2 Symmetrinen mitta

Mitaksi $e_m(u_m, d_m)$ kaavaan (2) voidaan valita esimerkiksi suosittu vektorien välisen kulman kosini [2], [3]

$$e_m(u_m, d_m) = 1 - \cos(u_m, d_m) = 1 - \frac{u_m \cdot d_m}{|u_m||d_m|} \tag{6}$$

tai p-normi

$$e_m(u_m, d_m) = a(p)||u_m - d_m||_p = a(p)\left(\sum_k |u_{mk} - d_{mk}|^p\right)^{1/p} \quad , \tag{7}$$

missä

$$a(p) = \begin{cases} \left(\frac{1}{1^p+1^p}\right)^{1/p} = \left(\frac{1}{2}\right)^{1/p} & \text{, kun } p < \infty \\ 1 & \text{, kun } p = \infty \end{cases}$$

palauttaa etäisyydet välille $[0, 1]$.

Kaikkein yksinkertaisinta olisi valita p-normi ja joko $p = 1$ tai $p = \infty$, koska kosinimitassa ja muilla p:n arvoilla joudutaan sekä korottamaan potenssiin että ottamaan juuria. Jos $p = 1$, mitaksi saadaan

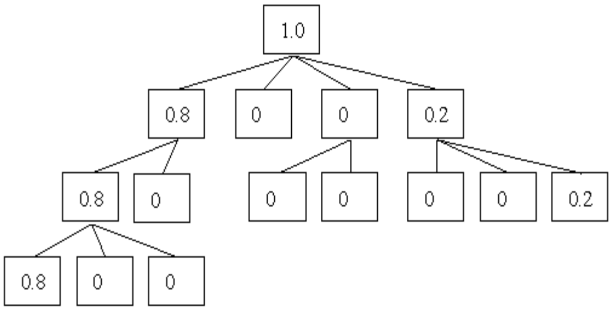
$$e_m(u_m, d_m) = \frac{1}{2}\sum_{m=1}^{N} w_m \left(\sum_k |u_{mk} - d_{mk}|\right) \quad . \tag{8}$$

Jos taas $p = \infty$,

$$e_m(u_m, d_m) = \sum_{m=1}^{N} w_m \left(\max_k |u_{mk} - d_{mk}|\right) \quad . \tag{9}$$

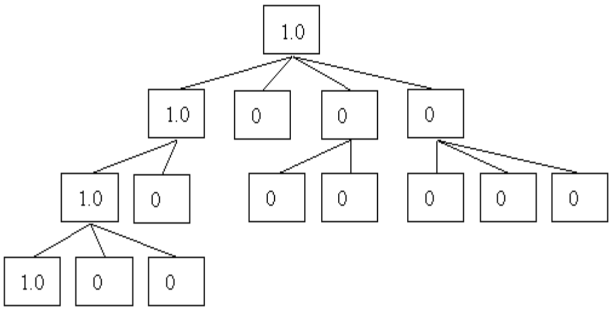
5.2.1 Esimerkki

Tarkastellaan muutamia erilaista dokumentteja ja erästä selkeää käyttäjäprofiilia ja tutkitaan, millaisia etäisyyksiä edellä mainituilla etäisyysmitoilla saadaan. Olkoot käyttäjän painokertoimet jakautuneet kuvan 4 mukaan.

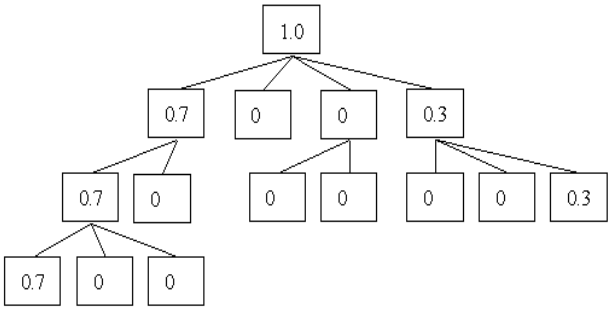


Kuva 4: Käyttäjäprofiili user1.

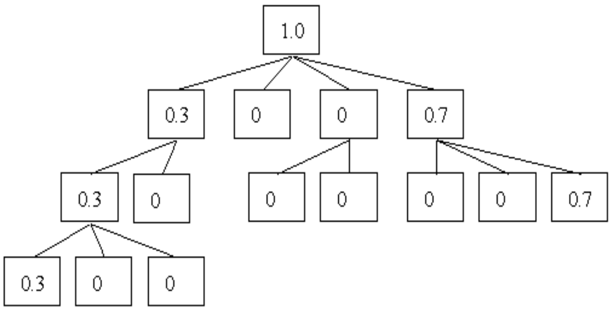
Verrataan viittä eri dokumenttia dok1 ... dok5 käyttäjäprofiiliin :



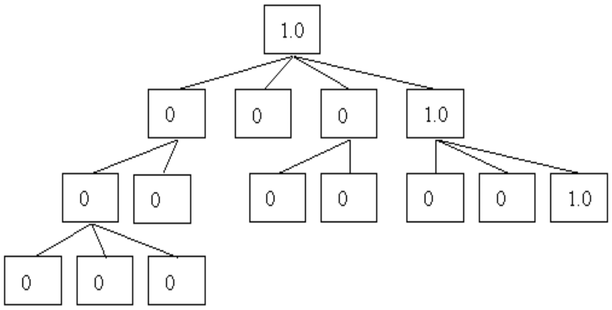
Kuva 5: Dokumentti dok1.



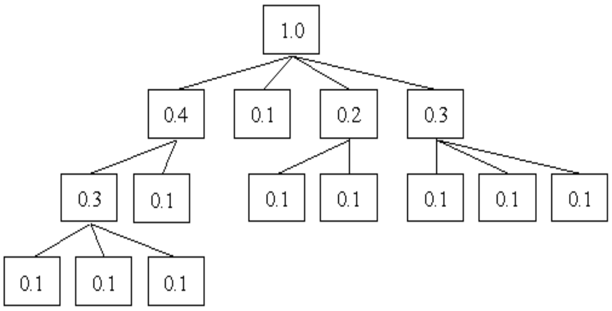
Kuva 6: Dokumentti dok2.



Kuva 7: Dokumentti dok3.



Kuva 8: Dokumentti dok4.



Kuva 9: Dokumentti dok5.

- dok1 osuu tarkasti aiheeseen, jossa profiilissa on suuri paino.
- dok4 osuu tarkasti aiheeseen, jossa profiilissa on pieni paino.
- dok2 on melko tarkasti profiilin muotoinen.
- dok3 on dok2:n peilikuva : suuret painot osuvat profiilin pieniin painoihin ja päinvastoin.
- dok5 liittyy yhtä paljon jokaiseen lehtitason käsitteeseen.

Kosinimitta		
dokumentti	$w_m = \frac{1}{N}$	$w_m = \left(\frac{1}{2}\right)^{N-m+1}$
dok1	0.0299	0.0299
dok2	0.0128	0.0128
dok3	0.3949	0.3949
dok4	0.7575	0.7575
dok5	0.3290	0.4009

p-normin mukainen mitta						
	$p = 1$		$p = 2$		$p = \infty$	
dokumentti	tasa	bias	tasa	bias	tasa	bias
dok1	0.2	0.2	0.2	0.2	0.2	0.2
dok2	0.1	0.1	0.1	0.1	0.1	0.1
dok3	0.5	0.5	0.5	0.5	0.5	0.5
dok4	0.8	0.8	0.8	0.8	0.8	0.8
dok5	0.6	0.65	0.4234	0.4522	0.5333	0.575

missä tasa : $w_m = \frac{1}{N}$ ja bias : $w_m = \left(\frac{1}{2}\right)^{N-m+1}$.

Sekä kosinimitalla että p-normimitalla dok2 arvioidaan osuvimmaksi dokumentiksi ja dok1 tulee toiseksi, koska molemmille mitoille etäisyys minimoituu vain kun $u = d$. Samasta syystä dok4 jää viimeiseksi. Kaikissa muissa tarkastelluissa tapauksissa paitsi 2-normissa dok3 voittaa dok5:n, koska dok5 osuu vain pienillä painoilla profiiliin ja menee monessa kohdassa ohi profiilista. Kun $p = 2$, painojen hajaantuminen useisiin eri solmuihin tuottaa dok5:lle pienemmän etäisyyden kuin dok3:lle.

5.2.2 Metriikan ehtojen toteutumisesta symmetriselle mitalle

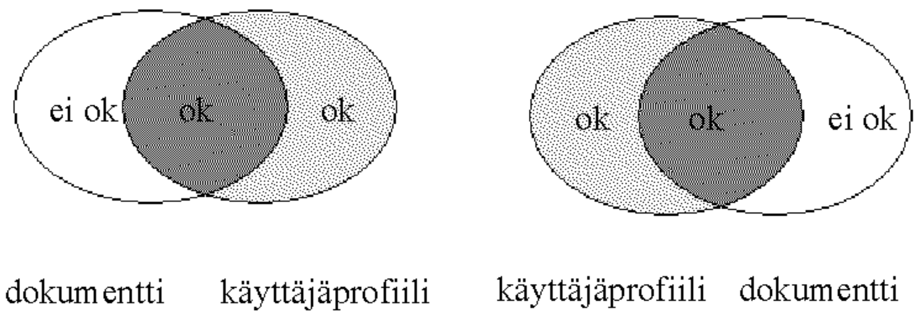
Tunnetusti p-normi toteuttaa metriikan ehdot. Kosinimitalle pätee yleisesti $e_m(u_m, d_m) = 0 \implies u_m = \alpha d_m$, mutta tässä sovelluksessa siitä seuraa

$$1 = \sum_k u_{mk} = \alpha \sum_k d_{mk} = \alpha \quad \implies \quad u_m = d_m \ .$$

Kosinimitalle metriikan ehdot (i) ja (ii) ovat voimassa, mutta (iii) ei.

5.3 Symmetrisyys

Vaikka käyttäjäprofiili ja dokumentin esitys ovat muodoltaan samanlaiset, jakaumalla esitetään erilaisia asioita käyttäjän ja dokumentin tapauksessa. Mitan ei ole siis tarkoituksenmukaista olla symmetrinen. Dokumentti on tyypillisesti muutamasta aiheesta ja paino keskittyy voimakkaasti yhteen tai kahteen komponenttiin. Sen sijaan sama käyttäjä voi olla suurin piirtein yhtä kiinnostunut useistakin asioista, eikä täydellinen osuma ole yleensä sellainen dokumentti, joka liittyy kaikkeen mahdolliseen, mistä käyttäjä on kiinnostunut.



Kuva 10: Hyvä osuma voi osua myös pelkästään käyttäjäprofiilin osapuuhun, joten ei ole välttämättä ideaalista vaatia $e_m(u_m, d_m) = 0 \implies u_m = d_m$

Kuvassa 10 näkyy tummalla alue, jossa sekä profiilissa että dokumentissa on painoa. Lisäksi harmaalla on merkitty alue, jossa profiilissa on painoa, mutta dokumentti ei osu siihen. Se on kuitenkin täysin hyväksyttävää eikä sen pitäisi kasvattaa etäisyyttä. Sen sijaan valkoinen alue merkitsee sellaisia dokumentin painoja, jotka eivät osu profiilin kiinnostuksiin, joten niiden painojen määrä kertoo, miten paljon dokumentti menee profiilista ohi (vastaa eräänlaista etäisyyttä). Jos käyttäjäprofiilin ja dokumentin järjestystä vaihdetaan etäisyysmitassa, pitäisi kuvan 10 mukaisesti tulla eri tulos.

Itse asiassa ei ollakaan kiinnostuneita siitä, miten paljon painot eroavat käyttäjäprofiilissa ja dokumentissa vaan kahdesta asiasta :

1. Kuinka hyvin dokumentti osuu käyttäjäprofiiliin ? (eikä toisinpäin)

2. Kuinka paljon painoa profiilissa on osumasolmuissa ?

Pilkotaan seuraavaksi ongelma osiin ja muodostetaan ensin mitta, joka mittaa kuinka hyvin dokumentti osuu profiiliin (nk. karkea mitta).

6 Karkea epäsymmetrinen etäisyysmitta

Muodostetaan mitta sille, miten hyvin dokumentin painojakauma osuu profiilin osapuihin, eikä verrata käyttäjäprofiilia kokonaisuudessaan dokumentin metadataan. Etäisyys muodostuu kuvan 10 valkoisen alueen painoista.

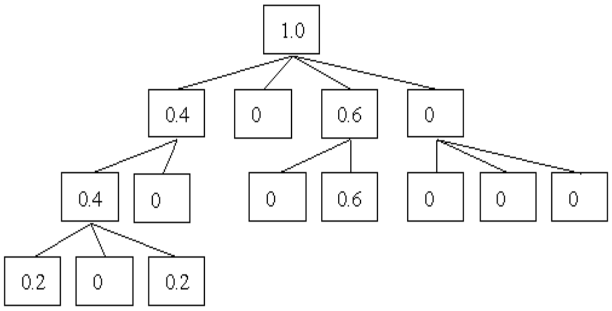
Merkitään symbolilla d_{mh} niitä dokumentin painoja, jotka eivät osu käyttäjäprofiiliin (hutipainot). Siis vastaavassa solmussa on käyttäjäprofiilissa painona u_{mh} nolla. Olkoon u käyttäjäprofiili ja d dokumentin esitys. Määritellään karkea mitta :

$$e(u, d) = \sum_{m=1}^{N} w_m \left(\sum_h |d_{mh}|^p \right)^{1/p} \tag{10}$$

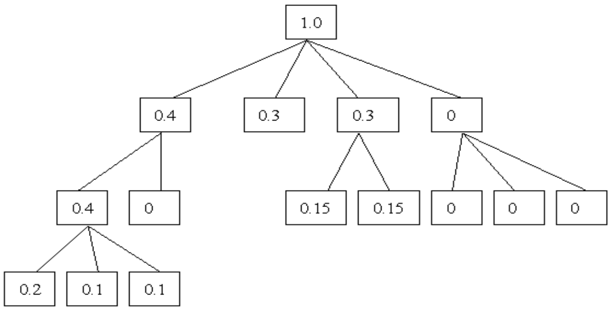
Tällä mitalla ei päästä vielä vertaamaan, miten hyvin eri dokumentit vastaavat profiilia vaan pikemminkin karsitaan pois sellaisia, jotka eivät osu profiiliin. Eri tasojen painottaminen painoilla w_m voidaan tehdä kuten luvussa 5.1.

6.1 Esimerkki

Luvussa 5.2.1 esitellyistä dokumenteista suurin osa osui koko painollaan käyttäjäprofiiliin user1. Otetaan vertailun vuoksi mukaan myös sellaisia dokumentteja, jotka eivät osu yhtä hyvin.

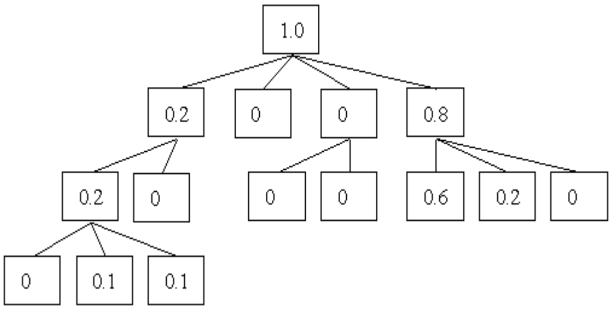


Kuva 11: Dokumentti dok6a.

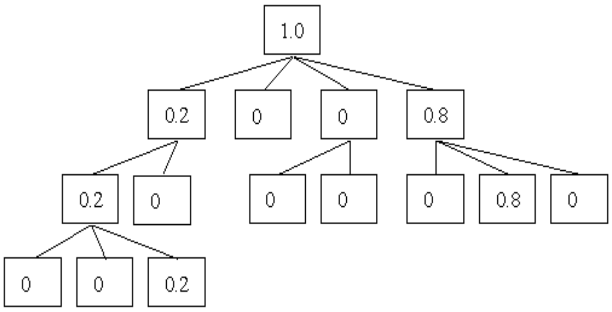


Kuva 12: Dokumentti dok6b.

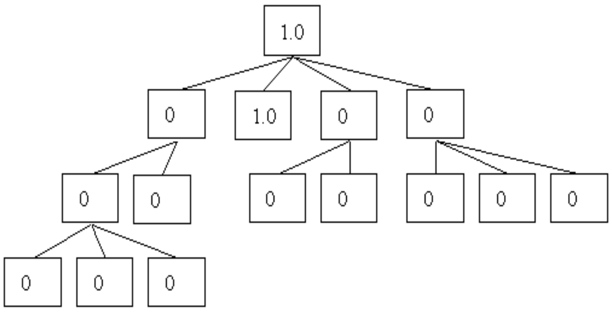
Dokumentit 6a ja 6b ovat toistensa parit siinä mielessä, että kaikki profiiliin user1 liittyvät painot ovat dokumenteissa samat. Näiden dokumenttien pitäisi siis osua täsmälleen yhtä hyvin eli saada sama etäisyysmitan arvo. Dokumentit 7a & 7b ja 8a & 8b ovat samalla tavoin parit.



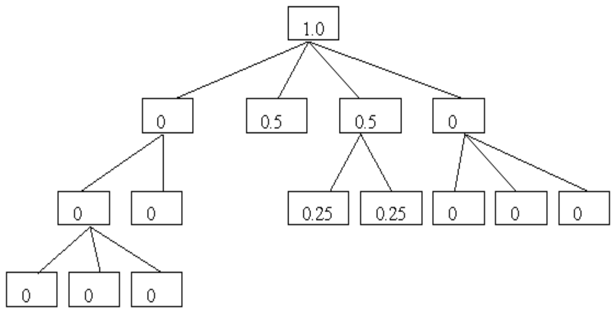
Kuva 13: Dokumentti dok7a.



Kuva 14: Dokumentti dok7b.



Kuva 15: Dokumentti dok8a.



Kuva 16: Dokumentti dok8b.

- dok6a & dok6b osuvat profiiliin pienellä painolla alas lehtitasolle asti.
- dok7a & dok7b osuvat hyvin ylemmillä tasoilla, mutta eivät enää ollenkaan lehtitasolla.
- dok8a & dok8b eivät osu profiiliin ollenkaan, joten etäisyyden pitäisi olla yksi.

Karkea mitta						
	$p = 1$		$p = 2$		$p = \infty$	
dokumentti	tasa	bias	tasa	bias	tasa	bias
dok1	0	0	0	0	0	0
dok2	0	0	0	0	0	0
dok3	0	0	0	0	0	0
dok4	0	0	0	0	0	0
dok5	0.5667	0.625	0.2505	0.2586	0.1333	0.125
dok6a	0.6667	0.7	0.6108	0.6162	0.6	0.6
dok6b	0.6667	0.7	0.3951	0.3948	0.3	0.3
dok7a	0.6	0.7	0.4268	0.4822	0.4	0.45
dok7b	0.6	0.7	0.5415	0.6123	0.5333	0.6
dok8a	1	1	1	1	1	1
dok8b	1	1	0.6440	0.6361	0.5	0.5

Koska sekä profiilit että dokumenttien metadata on normeerattu tasoittain 1-normin mielessä, se antaa etäisyysmittana kaikkein järkevimpiä tuloksia. Sekä 2-normin että maksiminormin tapauksessa etäisyyteen vaikuttaa se, miten paljon painot ovat jakautuneet eri haaroihin puussa, riippumatta siitä, osuvatko ne käyttäjäprofiiliin, millä ei pitäisi olla mitään merkitystä. Vain 1-normi antaa saman etäisyyden pareille 6a & 6b, 7a & 7b ja 8a & 8b.

6.2 Metriikan ehtojen toteutumisesta karkealle mitalle

Metriikan ehdon (i) kohdat $e_m(u_m, d_m) \geq 0$ ja $u_m = d_m \implies e_m(u_m, d_m) = 0$ toteutuvat, mutta $e_m(u_m, d_m) = 0 \not\Rightarrow u_m = d_m$. Ehdot (ii) ja (iii) eivät toteudu.

Karkean mitan lisäksi pitää pystyä tekemään ero esim. dokumenttien dok1 ja dok4 välille, koska dok1 osuu selvästi suurempaan profiilin painoon.

7 Tarkennusmitta karkealle mitalle

Karkea mitta ottaa huomioon vasta sen, miten paljon dokumentti osuu profiiliin, mutta ei tee eroa sille, miten suuri paino profiilissa on osumasolmuissa. (Kuvan 10 tumma alue.) Tarkennukseksi voidaan laskea esimerkiksi dokumentin ja profiilin pistetulo tasoittain ja ottaa niistä painotettu summa. (Kyseessä ei ole vektorien välisen kulman kosini, koska niitä ei ole normeerattu 2-normin mielessä.) Eri tasoilla saadut tulokset voidaan painottaa jälleen joko tasan tai lehtisolmuista alkaen suuremmilla painoilla.

Pistetulo tasolla m :

$$sim(u_m, d_m) = \sum_k u_{mk} \cdot d_{mk} \quad , \tag{11}$$

mitä voidaan pitää samankaltaisuuden mittana. Siitä saadaan etäisyysmitta esimerkiksi näin :

$$f_m(u_m, d_m) = 1 - sim(u_m, d_m) = 1 - \sum_k u_{mk} \cdot d_{mk} \quad , \tag{12}$$

ja eri tasoilta saadut etäisyydet yhdistetään

$$F(u, d) = \sum_{m=1}^{N} w_m f_m(u_m, d_m) = \sum_{m=1}^{N} w_m \left(1 - \sum_k u_{mk} \cdot d_{mk} \right) \quad . \tag{13}$$

7.1 Esimerkki

Tarkennusmitta		
dokumentti	$w_m = \frac{1}{N}$	$w_m = \left(\frac{1}{2}\right)^{N-m+1}$
dok1	0.2	0.2
dok2	0.38	0.38
dok3	0.62	0.62
dok4	0.8	0.8
dok5	0.7533	0.79
dok6a	0.7333	0.76
dok6b	0.7333	0.76
dok7a	0.84	0.88
dok7b	0.84	0.88
dok8a	1	1
dok8b	1	1

Havaitaan siis, että tarkennusmitta löytää keskinäisen järjestyksen dokumenteille dok1 ... dok4. Lisäksi parit säilyvät, kuten pitikin.

7.2 Metriikan ehtojen toteutumisesta tarkennusmitalle

Tarkennusmitta on symmetrinen (ehto (ii)), mutta se ei haittaa, koska sitä käytetään vain osumasolmuissa (kuvan 10 tummassa alueessa) ja karkea mitta vaikuttaa muualla. Kuvan 10 valkoisessa alueessa profiilin painot ovat nollia ja harmaassa alueessa dokumentin painot ovat nollia, joten kaavan (11) termit ovat nollasta poikkeavia vain tummassa alueessa. Metriikan ehto (i) toteutuu muuten, paitsi $u_m = d_m \nRightarrow f_m(u_m, d_m) = 0$. Myös kolmioepäyhtälö (iii) toteutuu.

8 Mittojen yhdistäminen

Karkea mitta arvolla $p = 1$ antaa järkevän arvion siitä, osuuko dokumentti yleensä ollenkaan ja tarkennusmitta antaa lisäinformaatiota niistä dokumenteista, jotka osuvat (luvut 6 & 7). Mikä on näiden mittojen välinen yhteys ja voidaanko niitä käyttää jonkinlaisena yhdistelmänä ?

8.1 Karkean mitan ja tarkennusmitan välinen yhteys

Tarkastellaan tasolla m niitä dokumentin d painoja, jotka eivät mene ohi profiilista u. Merkitään niitä d_{mr} ja vastaavia profiilin painoja u_{mr}. Tällöin

$$\sum_r u_{mr} = 1 \qquad \forall m \ .$$

Karkea etäisyysmitta tasolla m on

$$e_m(u_m, d_m) = \sum_h d_{mh} = 1 - \sum_r d_{mr} \quad ,$$

ja kun kaikki tasot lasketaan mukaan, etäisyydeksi saadaan

$$E(u, d) = \sum_{m=1}^{N} w_m e_m \quad .$$

Tarkennusmitta tasolla m on

$$\begin{aligned} f_m(u_m, d_m) &= 1 - \sum_k u_{mk} d_{mk} = 1 - \sum_h u_{mh} d_{mh} - \sum_r u_{mr} d_{mr} \\ &= 1 - \sum_r u_{mr} d_{mr} \quad , \end{aligned}$$

koska $u_{mh} = 0 \quad \forall h$.

$$\begin{aligned} \sum_r u_{mr} d_{mr} \leq \sum_r u_{mr} (\max_r d_{mr}) &= (\max_r d_{mr}) \sum_r u_{mr} \\ &= \max_r d_{mr} \leq \sum_r d_{mr} \\ &= 1 - e_m(u_m, d_m) \quad , \end{aligned}$$

joten

$$
\begin{aligned}
f_m(u_m, d_m) &= 1 - \sum_r u_{mr} d_{mr} \\
&\geq 1 - (1 - e_m(u_m, d_m)) \\
&= e_m(u_m, d_m) \quad .
\end{aligned}
$$

Kun lasketaan tarkennusmitta eri tasoilta yhteen, saadaan

$$
F(u, d) = \sum_{m=1}^{N} w_m f_m(u_m, d_m) \geq \sum_{m=1}^{N} w_m e_m(u_m, d_m) = E(u, d) \quad .
$$

Siis dokumentin karkea etäisyysmitta antaa alarajan tarkennusmitan arvolle.

8.2 Yhdistetty mitta

Lähtökohtana pidetään karkeaa mittaa arvolla $p = 1$ ja kun se antaa tasatuloksen, järjestyksen määrää tarkennusmitta $f(u, d)$. Dokumenttien saama järjestys eri painotuksilla käy ilmi seuraavasta taulukosta.

Karkea mitta + tarkennus			
	$w_m = \frac{1}{N}$	$w_m = \left(\frac{1}{2}\right)^{N-m+1}$	$w_m = \left(\frac{1}{2}\right)^m$
I	dok1	dok1	dok1
II	dok2	dok2	dok2
III	dok3	dok3	dok3
IV	dok4	dok4	dok4
V	dok5	dok5	dok7a & dok7b
VI	dok7a & dok7b	dok6a & dok6b	dok5
VII	dok6a & dok6b	dok7a & dok7b	dok6a & dok6b
VIII	dok8a & dok8b	dok8a & dok8b	dok8a & dok8b

Dokumenttien dok6a ja dok7a välinen järjestys riippuu siitä, kumpi painotus karkeassa mitassa on valittu. Jos painot ovat tasajakautuneet, dok7a voittaa jo karkealla mitalla. Jos taas lehtisolmuja painotetaan, ovat dok6a ja dok7a yhtä hyviä karkealla mitalla ja dok6a voittaa tarkennuksessa. Ylempiä tasoja korostava painotus korottaa dokumenttiparin dok7a & dok7b liian korkealle.

Seuraavaksi muodostetaan sellainen yhdistetty mitta, jonka arvot asettuvat kullakin painokertoimen valinnalla edellä olevan taulukon mukaiseen järjestykseen.

Jaetaan väli $[0, 1]$ tasan n:ään yhtä pitkään väliin. Välin pituus on h , $h = \frac{1}{n}$. Pyöristetään karkea mitta h:n tarkkuudella. Tämä jakaa dokumentit $(n+1)$:een luokkaan, joiden sisäinen järjestys selvitetään tarkennusmitan avulla. Luokkien välinen järjestys säilytetään. Jos karkea mitta pyöristyy jo ykköseen, siihen ei lisätä enää tarkennusta. Matemaattisesti tämä voidaan ilmaista:

$$E(u, d) = h \left\lfloor \frac{1}{h} e(u, d) + 0.5 \right\rfloor + h f(u, d) \quad , \tag{14}$$

missä $\lfloor \cdot \rfloor$ on lattiafunktio, $e(u, d)$ on karkea mitta $(p = 1)$ ja $f(u, d)$ on tarkennusmitta.

Esimerkkejä

$n = 10$, $h = 0.1$				
	$w_m = \frac{1}{N}$		$w_m = \left(\frac{1}{2}\right)^{N-m+1}$	
I	dok1	0.02	dok1	0.02
II	dok2	0.038	dok2	0.038
III	dok3	0.062	dok3	0.062
IV	dok4	0.08	dok4	0.08
V	dok5	0.6753	dok5	0.679
VI	dok7a & dok7b	0.684	dok6a & dok6b	0.776
VII	dok6a & dok6b	0.7733	dok7a & dok7b	0.788
VIII	dok8a & dok8b	1	dok8a & dok8b	1

$n = 2,\ h = 0.5$				
	$w_m = \frac{1}{N}$		$w_m = \left(\frac{1}{2}\right)^{N-m+1}$	
I	dok1	0.1	dok1	0.1
II	dok2	0.19	dok2	0.19
III	dok3	0.31	dok3	0.31
IV	dok4	0.4	dok4	0.4
V	dok6a & dok6b	0.8667	dok6a & dok6b	0.88
VI	dok5	0.8767	dok5	0.895
VII	dok7a & dok7b	0.92	dok7a & dok7b	0.94
VIII	dok8a & dok8b	1	dok8a & dok8b	1

Tapaukset $n = 5$ ja $n = 8$ ovat näiden väliltä ja niihin liittyvät tulokset löytyvät liitteestä.(A.1)

Pyöristyksen poisjättäminen vastaa raja-arvoa, kun $h \to 0$, joka on pelkkä karkea mitta. Jos taas $h \to 1$, niin karkean mitan vaikutus vähenee ja pelkkä tarkennusmitta määrää lopulta koko mitan (ellei pyöristys ole jo niin karkea, että mitaksi on saatu 1).

9 Yhteenveto

Elektronisessa täsmäjakeluympäristössä pyritään henkilökohtaistamaan tiedonjakelua vastaamaan kunkin käyttäjän tarpeita. Käyttäjän kiinnostusta kuvataan käyttäjäprofiililla ja dokumentin ominaisuuksia metadata-esityksellä. Käyttäjää eniten kiinnostavien dokumenttien suodattamiseksi dokumenttijoukosta profiilia ja metadataesitystä verrataan toisiinsa. Profiilin ja dokumentin esityksen välinen etäisyysmitta määrää, kuinka hyvin dokumentti vastaa profiilia.

Tässä työssä tutkittiin, millaisia ominaisuuksia käyttäjäprofiilin ja dokumentin välisen mitan olisi mielekästä toteuttaa, jotta saadut tulokset vastaisivat mahdollisimman hyvin intuitiivista kuvaa siitä, millainen on hyvin osuva dokumentti. Samalla myös kehitettiin mitta, joka toteuttaa halutut ominaisuudet. (Luku 8.2)

Tärkeä havainto on huomata, että täydellinen dokumentti ei ole täsmälleen samanlainen kuin käyttäjäprofiili vaan koostuu kahdenlaisesta osumasta :

- Missä määrin dokumentti osuu käyttäjän kiinnostusaiheisiin.
- Kuinka suuri merkitys (paino) käyttäjäprofiilissa on dokumentin aiheelle.

Näin ollen kehitetty etäisyysmittakin koostuu kahdesta osasta (karkea mitta ja tarkennusmitta).

Jos esimerkiksi profiilin kiinnostusalueet olisivat numeerinen matematiikka (painolla 0.4), asuntojen myynti-ilmoitukset (0.3) sekä YTV:n bussi-aikataulut (0.3), on ilmeistä että eniten kiinnostavat dokumentit eivät ole profiilin näkoisiä vaan osuvat aina johonkin profiilin osaan. Lisäksi ei ole yhdentekevää verrataanko profiilia dokumenttiin vai dokumenttia profiiliin. Kaiken, mitä profiilissa on, ei tarvitse löytyä dokumentista, mutta kaiken, mitä dokumentissa on, olisi hyvä löytyä profiilista (epäsymmetrisyys).

Viitteet

[1] Bloedorn, E., Mani, I., MacMillan R.(1996). Representational Issues in Machine Learning of User Profiles,
AAAI Spring Symposium on Machine Learning in Information Access.

[2] Sheth, Beerud (1994). A Learning Approach to Personalized Information Filtering, M.I.T.

[3] Foltz, P.W., Dumais, S. (1992). Personalized Information Delivery : An Analysis of Information Filtering Methods.
Communications of the ACM 35(12)

[4] Savia, Eerika (1997). Adaptiivinen oppimismenetelmä käyttäjäprofiilien päivittämiseen. Matematiikan erikoistyö, TKK.

A Liite : Tuloksia

A.1 Profiili user1

Painottamalla puun ylempiä tasoja enemmän kuin alempia saadaan etäisyyksiksi

p-normin mukainen karkea mitta, $w_m = \left(\frac{1}{2}\right)^m$			
dokumentti	$p = 1$	$p = 2$	$p = \infty$
dok1	0	0	0
dok2	0	0	0
dok3	0	0	0
dok4	0	0	0
dok5	0.5	0.2438	0.15
dok6a	0.65	0.6081	0.6
dok6b	0.65	0.4024	0.3
dok7a	0.45	0.3201	0.3
dok7b	0.45	0.4062	0.4
dok8a	1	1	1
dok8b	1	0.6597	0.5

Tarkennusmitta	
dokumentti	$w_m = \left(\frac{1}{2}\right)^m$
dok1	0.2
dok2	0.38
dok3	0.62
dok4	0.8
dok5	0.72
dok6a	0.72
dok6b	0.72
dok7a	0.8
dok7b	0.8
dok8a	1
dok8b	1

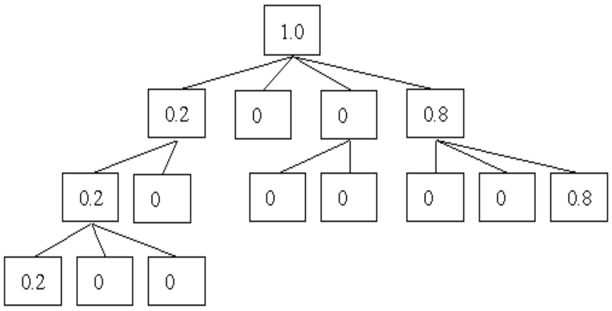
Yhdistetyn mitan arvoja eri jakovälin pituuksilla h:

$n=5,\ h=0.2,$ **user1**				
	$w_m = \frac{1}{N}$		$w_m = \left(\frac{1}{2}\right)^{N-m+1}$	
I	dok1	0.04	dok1	0.04
II	dok2	0.076	dok2	0.076
III	dok3	0.124	dok3	0.124
IV	dok4	0.16	dok4	0.16
V	dok6a & dok6b	0.7467	dok5	0.758
VI	dok5	0.7507	dok6a & dok6b	0.952
VII	dok7a & dok7b	0.768	dok7a & dok7b	0.976
VIII	dok8a & dok8b	1	dok8a & dok8b	1

$n=8,\ h=0.125,$ **user1**				
	$w_m = \frac{1}{N}$		$w_m = \left(\frac{1}{2}\right)^{N-m+1}$	
I	dok1	0.025	dok1	0.025
II	dok2	0.0475	dok2	0.0475
III	dok3	0.0775	dok3	0.0775
IV	dok4	0.1	dok4	0.1
V	dok6a & dok6b	0.7167	dok5	0.7238
VI	dok5	0.7192	dok6a & dok6b	0.845
VII	dok7a & dok7b	0.73	dok7a & dok7b	0.86
VIII	dok8a & dok8b	1	dok8a & dok8b	1

A.2 Toinen käyttäjäprofiili

Etäisyydet laskettiin vertailun vuoksi myös toiselle käyttäjäprofiilille, joka on muodoltaan profiilin user1 peilikuva.



Kuva 17: Käyttäjäprofiili user2.

Karkea mitta ottaa huomioon, missä solmuissa painoa on, mutta jättää huomiotta painojen suuruudet. Karkean mitan kannalta profiilit user1 ja user2 eivät eroa toisistaan ja niille saadaan täsmälleen samat tulokset.

Tarkennusmitta, user2		
dokumentti	$w_m = \frac{1}{N}$	$w_m = \left(\frac{1}{2}\right)^{N-m+1}$
dok1	0.8	0.8
dok2	0.62	0.62
dok3	0.38	0.38
dok4	0.2	0.2
dok5	0.8133	0.835
dok6a	0.9333	0.94
dok6b	0.9333	0.94
dok7a	0.76	0.82
dok7b	0.76	0.82
dok8a	1	1
dok8b	1	1

Yhdistetyn mitan arvoja profiilille user2

$n = 10,\ h = 0.1$, user2				
	$w_m = \frac{1}{N}$		$w_m = \left(\frac{1}{2}\right)^{N-m+1}$	
I	dok4	0.02	dok4	0.02
II	dok3	0.038	dok3	0.038
III	dok2	0.062	dok2	0.062
IV	dok1	0.08	dok1	0.08
V	dok7a & dok7b	0.676	dok5	0.6835
VI	dok5	0.6813	dok7a & dok7b	0.782
VII	dok6a & dok6b	0.7933	dok6a & dok6b	0.794
VIII	dok8a & dok8b	1	dok8a & dok8b	1

$n = 2,h = 0.5$, user2				
	$w_m = \frac{1}{N}$		$w_m = \left(\frac{1}{2}\right)^{N-m+1}$	
I	dok4	0.1	dok4	0.1
II	dok3	0.19	dok3	0.19
III	dok2	0.31	dok2	0.31
IV	dok1	0.4	dok1	0.4
V	dok7a & dok7b	0.88	dok7a & dok7b	0.91
VI	dok5	0.9067	dok5	0.9175
VII	dok6a & dok6b	0.9667	dok6a & dok6b	0.97
VIII	dok8a & dok8b	1	dok8a & dok8b	1

Havaitaan, että peilikuvadokumentit ovat vaihtaneet keskenään paikkaa: dok1 $\leftrightarrow$ dok4 ja dok2 $\leftrightarrow$ dok3.