Tiedonhakumenetelmät

Helena Ahonen-Myka
Kevät 2004, osa 2
Relevanssin käsite. Evaluointi.

Tässä osassa...

- relevanssin käsitteestä
- tiedonhaun evaluoinnista

Relevanssi

- **relevanssi** on keskeinen käsite tiedonhaussa, mutta sen määrittely on vaikeaa
- tiedonhaun tarkoituksena on löytää **relevanttia tietoa** sitä tarvitsevalle
  - Mitä relevanssi tarkoittaa?
  - Millainen tieto tai dokumentti on relevantti?
  - Kuka tekstin tai dokumentin relevanssin arvioi?
  - Millä kriteereillä?

Relevanssi

- hakutuloksia, indeksointia, ym. arvioidaan menetelmillä, jotka perustuvat relevanssin käsitteeseen
- käsitteen määrittelystä ei olla päästy yksimielisyyteen
  - yhteenkuuluvuus
  - vastaavuus
  - aiheenmukaisuus
  - hyödyllisyys
  - käyttökelpoisuus

Aihe- vs. käyttäjärelevanssi

- relevanssin määrittelyssä on kaksi pääsuuntaa:
  - aiherelevanssi: relevance to a subject, topicality, system relevance
    - pelkistetyimmillään sanojen täsmäyttämistä dokumenteissa ja kyselyissä
  - käyttäjärelevanssi: user relevance, user oriented view of relevance
    - perustuu käyttäjän arvioon dokumenttien käyttökelpoisuudesta

Aihe- vs. käyttäjärelevanssi

- perusoletus aiherelevanssista puhuttaessa: hakusanat (tai laajemmat kielen ilmaisut) voivat kuvata riittävällä tavalla dokumenttien ja hakutehtävien merkityksen
  - uskotaan, että parempi hakuavainten täsmäytys johtaa parempaan tulokseen
  - esim. tekstin merkitystä voidaan yrittää päätellä kehittyneillä lingvistisillä menetelmillä
  - lähelle täydellistä ei ole kumminkaan päästy

Aihe- vs. käyttäjärelevanssi

- aiherelevanssi on hyödyllinen käyttökelpoisuutensa takia (määriteltävyys, mitattavuus), mutta se ei kuvaa kaikkea relevanssiin liittyvää
- tutkimuksen päähuomio on siirtynyt käyttäjärelevanssin suuntaan

7

Eräs tarkempi luokittelu

- algoritminen relevanssi
  - kyselyn ja tekstin vastaavuus täsmäytysmenetelmän mukaan
- aiherelevanssi
  - kyselyn aiheen ja tekstin aiheen vastaavuus ihmisen tulkitsemana
- kognitiivinen relevanssi
  - dokumentin relevanssi tiedontarvitsijan tietämyksen tilan kannalta

8

Eräs tarkempi luokittelu (jatkuu)

- tilannerelevanssi
  - dokumentin relevanssi tiedontarvitsijan tilanteen, tehtävän tai ongelman kannalta
- motivaatio/tunnerelevanssi
  - dokumentin relevanssi tiedontarvitsijan tavoitteiden tai motiivien kannalta, esim. viihdearvo

9

Tiedonhaun evaluointi

- tiedonhakututkimuksessa voidaan yleensä arvioida menetelmiä (tai järjestelmiä) vain suhteessa muihin menetelmiin (tai järjestelmiin)
- tarkastelun kohteena on hakuprosessi
- haku = yhden hakutehtävän käsittelyyn kuuluva toiminta + hakutehtävän, kyselyn ja hakutuloksen muodostama tietokokonaisuus

10

Tiedonhaun evaluointi

- tiedonhaun tutkimuksessa vertaillaan tavallisesti eri menetelmiä koeasetelman avulla
- usein käytetään ns. tiedonhaun laboratoriomallia
  - valitaan dokumenttikokoelma (tietokanta) ja joukko hakutehtäviä
  - etsitään kullekin hakutehtävälle kaikki sen kannalta relevantit dokumentit → saantikanta
    - käytännössä tietokannat ovat usein niin laajoja, että tyydytään vain jonkinlaiseen otokseen relevanteista dokumenteista

11

Tiedonhaun evaluointi

- hakutehtävät ovat tyypillisesti aihehakuja
- relevanssiarviot ovat tyypillisesti binäärisiä
  - dokumentti joko on tai ei ole relevantti hakutehtävän kannalta
  - moniasteisiakin relevanssiarvioita voitaisiin käyttää (esim. olennainen / hyödyllinen / marginaalinen / epärelevantti)
- hakutehtävistä muodostetut testikyselyt täsmäytetään testattavilla menetelmillä tietokannan dokumentteihin
- hakutuloksia arvioidaan evaluointikriteerien avulla

12

Evaluointikriteerit

- tavallisimmat evaluointikriteerit
  - saanti (recall)
  - tarkkuus (precision)
  - tuloksen koko
  - vastausaika

13

Saanti ja tarkkuus

- hakutulos jakaa tietokannan dokumentit aina kahteen ryhmään
  - haussa löydetyt
  - haussa hylätyt
- periaatteessa kaikille tietokannan dokumenteille pitäisi tehdä relevanssiarvio, jolloin dokumentit voidaan jakaa
  - haun kannalta relevantteihin ja
  - haun kannalta epärelevantteihin

14

Saannin ja tarkkuuden määrittelytaulukko

	Rele	vanssi	arvio
Haun tulos	Relevantti	Epärelevantti	Summa
Löydetty	a osumat	b häly	a + b löydetyt
Hylätty	c unohdetut	d sivuutetut	c + d hylätyt
Summa	a + c relevantit	b + d epärelevantit	a + b + c + d tietokanta

15

Saanti ja tarkkuus

- saanti
  - hakutuloksen osumien suhde kaikkiin relevantteihin dokumentteihin: a / (a + c)
  - kuinka suuri osa tietokannan sisältämistä relevanteista dokumenteista löydettiin
- tarkkuus
  - hakutuloksen osumien suhde kaikkiin löydettyihin dokumentteihin: a / (a + b)
  - kuinka suuri osuus hakutuloksesta koostui relevanteista dokumenteista
- molemmat esitetään joko desimaalilukuna välillä [0,1] tai prosenttilukuna välillä 0...100%

16

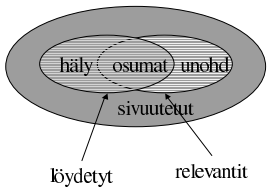
Saanti ja tarkkuus

- tarkkuus

$$\pi = \frac{osumat}{osumat + häly}$$

- saanti

$$\rho = \frac{osumat}{osumat + unohd.}$$



17

Saanti ja tarkkuus

- yhdessä saanti ja tarkkuus ovat tiedonhaun onnistuneisuuden konkreettisia mittareita
  - saanti kuvaa tiedontarvitsijan saaman tiedon määrää (suhteessa enintään saatavissa olevaan)
  - tarkkuus kuvaa sitä työtä, joka hänen on tehtävä erottaakseen hakutuloksen relevantit dokumentit

18

Saannin ja tarkkuuden suhde

- saannin ja tarkkuuden suhde on käänteinen
  - saannin parantaminen johtaa yleensä tarkkuuden huononemiseen ja päinvastoin
  - 100% saanti on aina saavutettavissa antamalla tulokseksi kaikki dokumentit → tarkkuus tällöin lähellä nollaa
- esim. jos kyselyyn lisätään hakutermejä, saanti kasvaa, mutta tarkkuus laskee
  - uudet hakutermit löytävät samasta asiasta eri sanoilla kirjoitetut dokumentit, mutta samalla nämä hakutermit voivat viitata myös täysin muihin aihepiireihin

19

Saannin ja tarkkuuden laskeminen

- allaoleva kuva esittää erästä tiedonhaun tulosta:
  - tuloksena on saatu 20 dokumenttia, jotka on numeroitu palautusjärjestyksessä
  - jokaisen dokumentin alla on dokumentin relevanssi tiedontarpeen suhteen (+ = relevantti, - = ei-relevantti)

d#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	-	-	-	+	-	+	-	-	-	-	-	+	-	-	-	+	-	-	+	-

20

Saannin ja tarkkuuden laskeminen

- oletetaan, että tiedetään tietokannassa olevan 10 tiedontarpeen kannalta relevanttia dokumenttia
- täystäsmäytys (esim. kyselynä Boolen lauseke)
  - hakutulos on joukko, jonka alkioille ei määritellä järjestystä
  - tarkkuus: 5/20 = 25%
  - saanti: 5/10 = 50%

21

Saannin ja tarkkuuden laskeminen

- osittaistäsmäytys (kyselynä joukko termejä)
  - hakutulos on lista, jonka alkioille hakujärjestelmä määrittelee relevanssilajittelujärjestyksen kyselyn ja dokumentin välisen täsmäävyyden mukaan
  - hakutuloksena on periaatteessa koko tietokanta todennäköisessä relevanssijärjestyksessä
    - kaikki relevantit dokumentit löydetään (jossain vaiheessa)
  - hakutuloksen hyvyyttä voidaan tarkastella vaiheittain

22

Saannin ja tarkkuuden laskeminen

Dokumentti #	Saanti %	Tarkkuus %
4	10	25
6	20	33
12	30	25
15	40	27
19	50	26

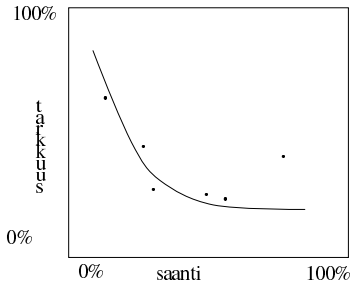
23

Saannin ja tarkkuuden esittäminen

- tavallisesti tarkkuusluvut huononevat saannin parantuessa
- yleensä tarkastellaan suurta joukkoa hakutuloksia ja ollaan kiinnostuneita keskimääräisistä saannin ja tarkkuuden arvoista
- esimerkiksi voidaan kerätä kunkin haun tarkkuusarvo, kun saanti on 10%, 20%,…, 100%, ja laskea keskimääräinen tarkkuus kullakin saannin tasolla
- keskiarvot voidaan esittää **saanti-tarkkuus -käyränä**

24

Saanti-tarkkuus -käyrä



25

d#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17...30...45
	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	+	+

	s%	t%		s%	t%
1:	0	0	17:	60	18
2:	20	50	...		
3:	20	33	30	80	13
4:	20	25	...		
5:	20	20	45	100	11
...					
9:	20	11			
10:	40	20			

26

Saanti-tarkkuus -käyrä

- saadaan tarkkuusarvot eri saantitasoille
  - saanti 20%, tarkkuus 50%
  - saanti 40%, tarkkuus 20%
  - saanti 60%, tarkkuus 18%
  - saanti 80%, tarkkuus 13%
  - saanti 100%, tarkkuus 11%
- pisteet sijoitetaan koordinaatistoon ja interpoloidaan käyrä pisteiden kautta
- yleensä siis käytetään 10% askelta ja arvot lasketaan usean haun keskiarvona

27

DCV-käyrä

- käyttäjät saattavat olla kiinnostuneita lähinnä vain ensimmäisistä dokumenteista
- voidaan tarkastella saantia ja tarkkuutta pisteissä, jotka vastaavat tiettyä vastausjoukon kokoa
  - 5 dokumentin jälkeen, 10 dokumentin jälkeen jne.
  - → DCV (Document Cut-off Value) -käyrä

28

d#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17...30...45
	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	+	+

	s%	t%	
2:	20	50	1. relevantti mukana
5:	20	20	
10:	40	20	2. relevantti
15:	40	13	
20:	60	15	3. relevantti
25:	60	12	
30:	80	13	4. relevantti
35:	80	11	
40:	80	10	
45:	100	11	5. relevantti

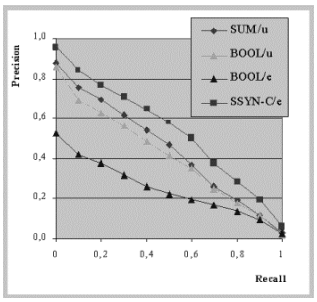
29

Hakumenetelmien vertailu

- hakutuloksista voidaan laskea kullekin haulle sen tuloksellisuus esim. saanti-tarkkuus –käyränä
- laskemalla keskiarvokäyrä koko hakujoukolle saadaan selville tutkitun tiedonhakumenetelmän suorituskyky
- tavallisesti tutkitaan usean erilaisen menetelmän keskinäistä suorituskykyä

30

Hakumenetelmien tuloskäyriä



31

Hakumenetelmien vertailu

- edellisen kalvon kuvassa esitetään neljän menetelmän suorituskykyä
  - kutakin menetelmää edustaa eri värinen saanti-tarkkuus –käyrä
  - kukin käyrä esittää yhden hakumenetelmän keskimääräistä suorituskykyä 30 haun joukossa saantitasoittain
- parhaan menetelmän keskimäärinen tarkkuus 50% saantitasolla on lähes 60% ja huonoimman noin 20% → tuloksellisuudessa näyttäisi olevan eroja

32

Hakumenetelmien vertailu

- tiedonhakumenetelmien kehittämisen kannalta on tärkeää arvioida, millaiset erot ovat olennaisia
- usein lasketaan suorituskäyrän keskiarvo 11 mittauspisteen avulla
  - keskiarvo tarkkuusarvoista saantitasoilla 0-100% (kymmenen prosentin välein)
  - esimerkissä parhaan menetelmän suorituksen tarkkuuskeskiarvo yli saantitasojen on noin 60%, muiden noin 50%, 40% ja 20%

33

Hakumenetelmien vertailu

- erojen merkitys käytännön kannalta
  - ero yli 15%: olennainen
  - ero 10-15%: merkittävä
  - ero 5-10%: kiinnostava
  - ero alle 5%: marginaalinen
- lisäksi lasketaan tilastollinen merkitsevyys
  - Kuinka todennäköistä on, että kyseinen ero on voinut syntyä sattumalta?
  - tilastotestit, mm. t-testi

34

Hakumenetelmien vertailu

- tulosten tulkinta saanti-tarkkuus –käyrillä voi olla ongelmallista, jos eri hakutehtävien saantikannat (=tietokannassa todella olevien relevanttien dokumenttien lukumäärät) vaihtelevat paljon
- tieto siitä, että paras hakumenetelmä saavuttaa 60% saantitasolla 50% tarkkuuden, ei kerro, montako dokumenttia hakija saa
- saantikantojen kokojen vaihtelu on ongelma myös DCV-käyrillä
  - jos saantikannan koko on 5 dokumenttia, ei tarkkuus tuloksen koolla 50 voi olla korkea

35

Tässä osassa käsiteltiin

- erilaisia näkökulmia relevanssin käsitteen määrittelyyn
- tiedonhakumenetelmien ja –järjestelmien evaluoinnin perusperiaatteita
  - evaluointikriteerit saanti ja tarkkuus
  - yhden haun tuloksen evaluointi
  - yhden menetelmän tuloksellisuuden tarkastelu hakujoukon avulla
  - usean menetelmän vertailu

36