Tiedonhaun perusteet tentti 10.12.2018 klo 14-16. Poistua saa 14:30->

Kaikki tehtävät arvostellaan pistein 0 – 5. Pistemäärä voi olla desimaaliluku. Yritä parhaasi ja onnea tenttiin!

1. Muun muassa Googlen hakutulosten järjestäminen perustuu Page Rank -algoritmiin. Kuvaa Page Rank -algoritmin toimintaperiaate. [ohje: essee, saa käyttää kuvallista kerrontaa]. (5p)

2. Vastaa lyhyesti, max. 30 sanaa per kohta (5p)
   a. Karsinta eli stemmaus.
   b. Sivun ulkoinen hakukoneoptimointi
   c. Haun kannalta relevantteja dokumentteja on tietokannassa 6. Hakutulos on 4 dokumenttia, joista haun kannalta relevantteja on 3 ja epärelevantteja 1. Mitä ovat haun saanti, tarkkus ja kumuloituva hyöty (CG@4, kun relevantin arvo on 1 ja epärelevantin 0)?
   d. Web crawling
   e. Käyttäjärelevanssi

3. Tiedonhaun suhde tiedonhankintaan? [ohje: essee, saa käyttää kuvallista kerrontaa] (5p)

4. Muotoile (yksi) ***kysely*** (hakulauseke) aiheesta *"tiedonhakumallit ja hakukoneet semanttisessa internetissä"*.* Järjestelmässä, johon kysely syötetään on käytössä Boolen operaattorit AND, OR ja NOT, sekä sulut (). Täsmäytysmenetelmä on täystäsmäytys. Lisäksi käytössä on fraasihaku (esim: "sana pari") ja katkaisuoperaattori * (esim: jäni*). Dokumenttien sanat on indeksoitu käsittelemättömänä (taivutusmuodot säilyttävä, ei stemmattu/lemmattu ja yhdyssanoja osittamaton indeksi).

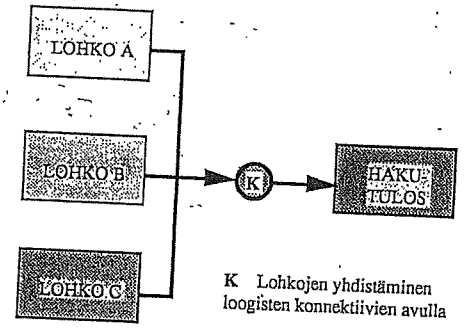
   Dokumentit ovat suomenkielisiä tekstidokumentteja, joita on yhteensä miljoonia ja ne käsittelevät erilaisia aiheita. Muuta ei tiedetä. Tuloksissa pyritään mahdollisimman suureen saantiin ja lisäksi hyvään tarkkuuteen, joten kyselyn muotoilussa on suositeltavaa soveltaa luennolla ja harjoituksissa esiteltyä ns. (yhdistettyä) lohkohakustrategiaa (periaate esitelty myös kääntöpuolella, jos ei muistu mieleen).

   Tehtävään ei ole olemassa täydellistä vastausta (eikä toisaalta helposti täysin vääräkään). Arvostelu perustuu käytetyn termistön osuvuuteen ja kattavuuteen, sekä hakulausekkeen järkevään rakenteeseen. Pyrin tulkitsemaan ratkaisujasi etujesi mukaisesti ;) (5p)

**)Tiedontarpeen tarkempi kuvaus: Halutaan dokumentit aiheesta tiedonhakumallit ja hakukoneet semanttisessa internetissä. Relevanteissa dokumenteissa kerrotaan, mitä ja millaisia hakumenetelmiä ja hakukoneita kehitetään ja käytetään semanttisen internetin yhteydessä."*

10 HAKUSTRATEGIAT

Sanaa hakustrategia on käytetty monessa mielessä. Jotkut käyttävät sitä tarkoittamaan hakuohjelmalle annettuja komentoja eli kyselyä (myös hakuprofiili, search formulation, search profile). Toiset tarkoittavat hakustrategialla haun kuluessa tehtäviä päätöksiä siitä, mikä on paras tapa jatkaa hakua eli *hakutaktiikkaa* (*search tactic, search heuristic*, katso Harter, 1986). Tässä kirjassa hakustrategia ja hakutaktiikka -termejä käytetään vastaavassa merkityksessä kuin sodankäynnissä : *Hakustrategia* on kokonaissuunnitelma tai lähestymistapa haun suorittamiseen ja *hakutaktiikkaa* ovat askeleet, jotka suoritetaan valitun strategian edistämiseksi. Yhdellä haulla voi olla yksi strategia, mutta siinä voidaan käyttää monia taktiikoita. Hakustrategiatyyppejä ovat : pikahaku, lohkohaku, helmenkasvatushaku, lohkojen peräkkäishaku ja selailuhaku (Harter, 1986, Mark Pejtersen, 1979). Hakutaktiikkaa tarkastellaan kohdassa 14.3.



Kuva 10.2. Lohkohaku

10.2 Lohkohaku

Lohkohaku (*building blocks strategy, analytic strategy*) on usein käyttökelpoisin hakustrategiatyyppi, jos pyritään kyselyyn, jonka saantia ja tarkkuutta sekä tuloksen kokoa voidaan säädellä vuorovaikutuksessa hakujärjestelmän kanssa. Lohkohakua käytettiin edellä hakuaiheen käsiteanalyysissa. Sen ydin on seuraava (Kuva 10.2) :

- tunnistetaan hakuaiheen keskeiset käsitteet (aspektit) ja niiden loogiset suhteet ;
- tunnistetaan kattavasti ne hakuavaimet, jotka edustavat kutakin käsitettä : sanat, fraasit, luokkasymbolit, koodit (esimerkiksi kielikoodit); yhtä käsitettä edustavat hakuavaimet muodostavat lohkon ;
- haetaan kullekin lohkolle hakutulos, jossa edellytetään, että tuloksen jokainen dokumentti sisältää ainakin yhden käsitteelle annetuista vaihtoehtoisista avaimista (haetaan siis lohkon avainten disjunktio, katso kohta 7.1) ; lohkot
- saadut ~~hakutulokset~~ liitetään toisiinsa siten kuin käsitteellisen hakusuunnitelman rajaavat ja rinnakkaiset suhteet edellyttävät.

Lohkohaussa hakutuloksen dokumenttien täytyy sisältää ainakin yksi hakuavain kustakin rajaavasta lohkosta. Lohkohakua on helppo ohjata laajempaan tai suppeampaan tulokseen karsimalla tai laajentamalla lohkojen sisältämiä hakuavaimia.