NTNU
Norges teknisk-naturvitenskapelige universitet

Fakultet for informasjonsteknologi, matematikk og elektroteknikk

Institutt for datateknikk og informasjonsvitenskap

Faglig kontakt under eksamen:
Institutt for datateknikk og informasjonsvitenskap
Heri Ramampiaro, 73593440

EKSAMEN I EMNE IT2801 INFORMASJONSGJENFINNING

Torsdag 24. mai 2007.
Tid: kl 09.00 – 13.00 (4 timer)

LØSNINGSFORSLAG

Hjelpemidler: D – Ingen trykte eller håndskrevne tillatt. **Kun** typegodkjent kalkulator er tillatt.

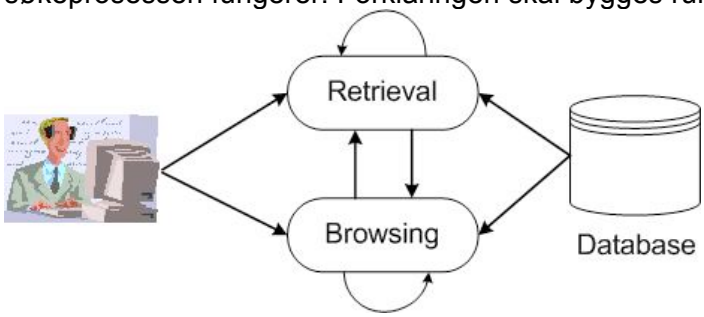
Sensuren faller: **14. Mai 2007**.

Svar **kort og konsist** på alle spørsmålene. **Stikkord** foretrekkes fremfor lange forklaringer.

Oppgave I (25%)

1. Forklar hvordan informasjonsgjenfinnings prosessen er bygd opp. Du må her tegne et blokkdiagram (med firkanter og piler) av denne prosessen. Tips: Dette er ikke tekstoperasjoner. (5%)

   Svar:
   Her forventes at studentene tegner opp en blokkdiagram som forklarer hvordan IR-søkeprosessen fungerer. Forklaringen skal bygges rundt brukerens informasjonsbehov.



2. Hvorfor er "***index terms***" viktig i informasjonsgjenfinningssammenheng? Hva er de viktigste kriteriene for valg indeks termer. Forklar. (4%)

   Svar:
   Index termenes rolle informasjonssammenheng er den skal representere dokumentene på en best mulig måte. Dette er hovedgrunnen til at de er viktige. Kriterier for valg av Index terms er: de skal være dekkene for innholdet i dokumentene de skal representere, med høyst mulig

diskrimineringsgrad, og de skal bidra til plassbesparelse.

3. Gjør rede for hvilke tekstoperasjoner man bør kjøre før selve søket av informasjon kan finne sted. Hva er hensiktene med disse operasjonene? (5%)

   Svar:
   Her skal man liste opp de fem tekstoperasjonene; leskikal analyse, fjerning av toppord, stemming, thesaurus bygging, og valg av indekstermer. De skal forklares. Hensiktene med dem er implisitt i forklaringene.

4. Hva er hovedforskjellene mellom "***Suffix arrays***" og "***Suffix tree***"? Forklar kort prinsippene bak dem. Du må bruke eksempler/illustrasjoner til å støtte forklaringen din. (5%)

   Svar:
   Som navnet sier er hovedforskjellene at det ene er tre det andre er en liste/array. I array'et listes opp kun der termene som skal indekseres dukker opp (occurences), mens i treet skal førstebokstaven i hver term være med sammen med occurences. På grunn av dette tar array'et mindre plass enn treet.

   Prinsippet med begge indekseringsmetodene er basert på at en tekst sees på som en lang streng, hvor hver ordposisjon i teksten er suffiks av strengen. Suffikstreet og suffiks arrayet er bygd basert på alle suffiksene i treet med startposisjon av suffiksene som "occurences". Hver posisjon peker på hver suffiks. Figurene nedenfor forklarer prinsippene.

5. "Inverted Files" er en indekseringsmetode. Beskriv hvordan du ville brukt "***inverted files***" til å indeksere dokumenter. Bruk gjerne et eksempel i forklaringen din. (6%)

   Svar:
   Her forventes at studentene forklarer prinsippet.

→

Oppgave II (25%)

1. Forklar hva som er hovedforskjellene mellom ***multimedia*** og ***tekst*** gjenfinning. (3%)

   Svar:
   Tekst er undersett av multimedia. Derfor er multimedia mer kompleks enn tekst. Av den grunn er multimedia gjenfinnings mer komplekst enn for tekst. Gjenfinning av tekst kan gjøres ved å bruke utvalgte termer som basis av indekser (dvs. indekstermer) men for multimedia må man velge egenskaper eller såkalte "features" som basis for indeksering. Features varierer avhengig av hvilke type media man skal gjenfinne.

2. Lag en ***taxonomy*** over multimedia datamodellen. Tips: Modellen er delt opp i flere lag. Det forventes at du gir minst et eksempel på multimedia objekt type for hvert lag. (6%)

   Svar:
   De forventes at studentene skal kunne gjengi i mest mulig grad følgende figur:

3. Hva er hensiktene med "features extraction"? Gi eksempler på features dersom det er snakk om (a) Bildegjenfinning og (b) Lydgjenfinning. (5%)

   Svar:
   Hensikten er hente ut egenskaper som best mulig representerer media som igjen gjør søke mest effektiv. (a) Bilde: fargehistogram, shape, teksturer, osv. (2) Audio: zero-crossing rate, silence ratio, spectrogram, energifordeling, etc.

4. Audio kan klassifiseres til tale og musikk.
   a. Vis hvordan du kan bruke forskjellene mellom disse to til å klassifisere audio ved hjelp av steg for steg klassifikasjon (step-by-step classification). Tegn opp et flytdiagram som viser stegene

(4%)

Svar:
Studentene skal vise og forklare følgende figur:

b. Er det mulig å skille to musikk typer (sjanger) basert på frekvensspekteret (frequency spectrum)? Begrunn svaret ditt. (3%)

Svar:
Ja, til en viss grad. Techno vil kunne vise mye ”pitch” mens klassisk musikk vil innholde lite pitch men mest harmonisk.

5. List opp hvilke aktuelle features som kan brukes til gjenfinning av video informasjon. Hva er hensiktene med ”segmentation” (segmentering) i video gjenfinning. (4%)

Svar:
Video features kan bruke mye av featurene for bilder i 3. men i tillegg må vi har særegne features som bevegelses informasjon/kamera bevegelser, r-frames, osv. Hensikten med ”segementering” er å dele video i shots som er naturlig å trekke features fra som igjen gjør det enklere å gjenfinne.

Oppgave III (30%)

1. Sammenlikn propablistisk og vektorbasert modellen. Hvilken modell ville du foretrekke dersom du skulle lage en tekstgjenfinningssystem. Du må her forklare styrke og svakheter med hver av modellene. (6%)

Svar:

Vektor modellen	Prablistiske modellen
Fordeler: - Delvis søk tillatt (bøhever ikke å finne eksakt match) - Rangering av resultater	- Delvis match tillatt - Rangering av søkeresultater basert på estimering av relevanssannsynlighet som basis

- Veldig enkel	
Ulemper: - Antar at alle termer er uavhengig - Kan ta med mange dokumenter som brukeren ikke mener er relevante	- Må estimere den initielle sannsynligheten noe som ikke alltid er rettfrem - Tar ikke hensyn til TF og IDF

2. ***Precision*** og ***recall*** er ofte to standard og mest brukte mål for evaluering et informasjons-gjenfinningssystem (IR-system). Hva er ulempen med å bruke disse målene? For å få et bedre resultat er det blitt foreslått et **brukerorientert mål.** Hvilke 4 begrep er relevante her. Forklar. (5%)

   Svar:
   Ulempen med precision og recall er ofte at de ikke tar brukeren i betraktning. Brukerorientert mål involverer følgende begreper:

   Det er viktig at "**novelity**" (deler av relevante dokumenter som er hentet men som brukeren ikke visste noe om på forhånd) og "**coverage**" (deler av dokumenter som kjent til å være relevante) blir nevnt og forklart.

3. Forklar kort prinsippene med User Relevant Feedback. Bruk gjerne figur til å støtte forklaringen din. (4%)

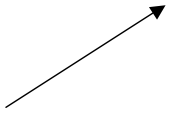
   Svar:
   URF kan forklares basert på figuren nedenfor:

4. Hva er hovedforskjellene mellom ”automatic local analysis” og ”automatic global analysis”? Fortell først i hvilken sammenheng disse begrepene brukes. (4%)

   Svar:
   ALA og **AGA** brukes i forbindelse med søkeforbedring/søkeutvidelse. Begge brukes til thesaurus bygging. Hovedforskjellene mellom disse er at ALA bygger thesaurus basert på uthentede (gjenfunnet) dokumenter mens AGA bygger thesaurus basert på hele dokumentsamlingen.

5. Konstruer en signatur fil av følgende tekst. Anta at du kan dele teksten i 4 blokker. Du må ellers gjøre dine antakelser angående **signatur funksjon**, og liknende.(6%)
   ”A tornado can cause sever destructions. Twister is a movie about a tornado”

   Svar:



6. Forklar prinsippet bak en metode for term revekting og utvideles av spørring (term reweighting and query expansion) for vektorbasert modellen. Tips: ”***Standard_Rochio***”. (5%)

   Svar:
   Standard Rochio formelen er som følger:

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \frac{\gamma}{|D_n|}\sum_{\vec{d}_j \in D_n}\vec{d}_j$$

   Formelen regner ut vektoren for en forbedret/utvidet spørring basert på den opprinnelige spørringen, de dokumentene som er hentet så langt som er relevante og de som ikke er relevante.

Oppgave IV (20%)

Svar ***rett/galt*** med ***begrunnelse*** på følgende utsagn. Hvert **riktig** og **begrunnet** svar belønnes med **2** poeng. **Feilsvar** får **-1,5** poeng. Mens **ubegrunnet** eller **ingen svar** er **0** poeng.

1. Fargehistogrammer er ikke nyttig for gjenfinning av videosnutter.
   (Rett/Galt)

   Svar:
   GALT. Kan se hver ramme i video som bilde og derfor kan fargehistogrammer benyttes som feature. Dessuten de brukes til skille shots fra hverandre (dvs videosegementering).

2. ”E-measure” kan sees som en generalisering av ”Harmonic Means”.
(Rett/Galt)

   Svar:
   RETT. E-measure reduseres til Harmonic Means nå konstanten beta i fomelen er lik 1.

3. Søkemotorer med ”Harvest” arkitektur er en variant av sentralisert web-søkemotor arkitektur.
(Rett/Galt)

   Svar:
   GALT. Dette er distribuert web-søkemotor.

4. Thesaurus er et verktøy til å forbedre spørringer og brukes ofte i forbindelse med automatic local analysis. (Rett/Galt)

   Svar:
   RETT. ALA bruker thesaurus til å utvide eller forbedre spørringer. Her brukes ofte ”co-occurence” informasjonen.

5. ”Gatherer” brukes som en ”crawler” i en distribuerte web-søkemotorer.
(Rett/Galt)

   Svar:
   RETT. Gatherer brukes i Harverst-baserte web-søkemotorer.

6. Bilde-features kan ikke brukes i video gjenfinningsprosesser.
(Rett/Galt)

   Svar:
   GALT. Kan se hver ramme i video som et bilde.

7. En SQL-database er godt egnet til innholdsbasert bildegjenfinning (content-based image retrieval).
(Rett/Galt)

   Svar:
   GALT. SQL database kan ikke håndtere eller lagre bilde-features.

8. Entropy er et konsept som brukes ofte til å karakterisere innholdet i informasjon. (Rett/Galt)

   Svar:
   RETT. Entropy forteller noe om mengden av informasjon i tekst.

9. I en Ordboksbasert (Dictionary-based) metode oppnås komprimering ved å bytte grupper av etterfølgende symboler (eller fraser) med peker til et innslag i ordboka.
(Rett/Galt)

   Svar:
   RETT. Det er dette som er prinsippet til ordboksbasert komprimeringsmetode som feks. LZ77 komprimeringsmetoden.

10. Micon er en viktig feature for bilder og brukes i bildegjenfinning, og kan sees som en analogi av r-frames innen video gjenfinning.
(Rett/Galt)

Svar:
GALT. Micon brukes som feature i videogjenfinning.