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
Robert Neumayer / Heri Ramampiaro, 73593440

EKSAMEN I EMNE IT2801 INFORMASJONSGJENFINNING

Fredag 22. mai 2009.
Tid: kl 09.00 – 13.00 (4 timer)

BOKMÅL

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

Sensuren faller: 12. juni 2009.

Denne oppgaven ble kvalitetssikret av: Prof. Kjetil Nørvåg

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

Oppgave I (25%)

1. Hva er hovedhensiktene med **informasjonsgjenfinningssystem**? Forklar hva som er hovedforskjellene mellom data- og informasjonsgjenfinning.

Svar: Hensiktene med IR-system er i hovedsakelig til å oppfylle brukerens informasjonsbehov ("users' information needs") og derfor tilby tilgang til informasjonen på en mest mulig brukervennlig måte.

Forskjellene mellom data- og informasjonsgjenfinning:

Data: strukturert, direkte matching, tolererer ikke feil i query.

Informasjon: ikke nødvendigvis strukturert – mest fritekst, delvis match og rangering av søkeresultater, tolerer feil i query.

2. Hva er rollen til indekstermer (index terms) i informasjonsgjenfinningssammenheng? Forklar hvordan "invertert indeks" (inverted index) fungerer. Bruk eksempel i forklaringen din.

Svar: Indekstermer: representere innholdet et dokument for å lette søke av relevant informasjon. Her kreves det forklaring av representasjonen av accurences av termer i dokumenter, posting lists, osv… Illustrasjonen er viktig for å få full pott.

3. Forklar hvordan boolske similaritetsmodellen (boolean similarity model) fungerer. Hva er grunnene til at boolske spørringer kan ha begrensede bruksområder?

Svar: Her er det viktig at studentene svarer:
- Beskriver boolsk modellen; bruken av boolsk operatorer, vektingsprinsipp (0 eller 1 avhengig om term er i et dokument eller ikke), matching mellom spørring og dokument…
- Grunnen til at boolsk har begrensede bruksområde: matchingsprinsippet (enten eller), ingen rangering, krav til brukeren når det gjelder kunnskap om boolsk operatorer og logisk tenking.

4. Drøft hovedutfordringene med web (dvs WWW) sett fra informasjonsgjenfinningsperspektiv.

Svar: Utfordringer: dynamisk egenskaper, kvaliteten på informasjonen (som påvirker relevans-begrepet), mengden av informasjon, etc.

5. Forklar hvordan et IR-system kan bli evaluert og hvordan dette kan gjøres.

Svar: Her er det meningen at studentene skal forklare prinsippet bak precision og recall, f-measure, R-precision, MAP. Det er viktig at studentene fokuserer på at evalueringen skal finne ut i hvilken grad et IR-system imøtekommer brukerens informasjonsbehov og evnen til å gjenfinne relevante dokumenter. Det gis bonus på om studentene nevner bruken av std. testkolleksjoner som TREC, GOV2, etc.

Oppgave II (25%)

1. Forklar stegene som er nødvendig fra man har en samling av dokumenter til disse er ferdig indeksert. (Tips: indexing pipeline).

Svar: Her er det nok at studentene forklarer de vannlige tekstoperasjonene: laxical analysis, stoppord fjerning, stemming, thesaurus bygging og valg av indekstermer.

2. Forklar hva begrepet "*edit distance*" er. Hva brukes det til? Hvis vi har en tekststrengen s_i, hvor $len(s_i)$ er lengden på s_i Vis at "*edit distance*" mellom s_1 og s_2 ikke kan være større enn max($len(s_1)$, $len(s_2)$).

Svar: Edit distance: Gitt to tekststreng s1 og s2, edit distance er antall operasjoner som trenges for å gjøre s1=s2, hvor en operasjon kan være sletting av tegn og/eller legge til et tegn. Jfr. defenisjonen kan ikke edit distance være større enn den lengste strengen.

3. "Feature" er et sentralt begrep i multimedia gjenfinning. Hva er hensiktene med "features"? Hvilke tre krav bør uthenting av features (*feature extraction*) oppfylle? Gi eksempler på features i forbindelse med videogjenfinning.

Svar: Feature er noe som best representer og/eller karakteriserer et media objekt ifm innholdsbasert gjenfinning av multimedia informasjon. Dette kan være farger, fargehistogram, form (shape), og texture for bilder. For video kan dette være R-frame, objekter, bevegelsesinformasjon, tekstannotering, osv.

Krav til feature extraction:
1. Det skal være komplett.
2. Det skal være effektivt i bruk (mest mulig diskrimineringseffekt);
3. Det skal være kompakt og ikke ta stor lagringsplass.

4. Forklar begrepet "Jaccard Coefficient". Hvorfor egner ikke denne seg så godt til rangering av søkeresultater?

Svar: Gitt to sett A og B, jaccard$(A,B) = |A \cap B| / |A \cup B|$
Det viktigste er at denne måler overlapp mellom to sett som kan i IR være to dokument med sett

av termer. I så måte ville den gi en indikasjon på likhet mellom dokumentene, men den tar ikke hensyn til termfrekvenser. Den trenger dessuten en mer sofistikert normalisering av lengder…

Oppgave III (30%)

1. Sammenlikn *sannsynlighetsmodellen* – **Okapi BM25** og *vektormodellen* – **TF/IDF**. Hvilken ville du foretrekke dersom du skulle lage en tekstgjenfinningssystem. Grunngi svaret ditt. (Tips: Fokuser på prinsippene, ulempene og fordelene).

Svar: Her skal det forklares Okapi BM25 og TF/IDF modellen. Det er viktig at studentene nevner at begge modellene bruker term frekvens (TF) i modellen og en form for IDF. BM25 tar i tillegg hensyn til dokumentlengden i vektingen av termene. Tester har vist at BM25 gir bedre resultater enn TF/IDF modellen.

2. Testkolleksjoner (Test Collections) brukes ofte i evaluering av informasjonsgjenfinningssystemer. Gi eksempler på eksisterende testkolleksjoner. Hvordan brukes de? Hva er *R-precision*, *F-measure* og *MAP* (*mean average precision*)?

Svar: Her gis poeng for beskrivelse av TREC, og GOV2 etc.
R-Precision: Presisjonsverdi når R antall dokumenter er fremhentet, hvor R er totalt antall relevante dokumenter i dok.kolleksjonen.
F-measure: 2PR/(P+R)
MAP: se side 147 i læreboka.

3. Utvidelse av spørringer (query extension).
    a. Forklar hva er hensiktene med utvidelse av spørringer (queries).

    Svar: For å forbedre eller utvide spørringer s.a man får bedre precision or recall.

    b. Forklar kort prinsippene med **Rocchios** metode for ***User Relevant Feedback*** (URF).

    Svar: Viktig her at studentene viser at de har forstått prinsippet. Det er en stor bonus hvis de klarer å gjengi formelen og illustrere bruken.

    c. Hva er hovedforskjellen mellom "*automatic local analysis*" og "*automatic global analysis*"?

    Svar: ALA vs AGA: begge brukes til å bygge thesaurus ifm URF, mens ALA bruker dokumentene fra søkeresultatet til dette, bruker AGA hele kolleksjonen.

4. Forklar forskjellige metoder for indeksering av tekstdokumenter i informasjonsgjenfinningssystemer.

Svar: Her kan studenten forklare prinsippet bak invertert indekser, suffix tre eller array, signatur fil.

5. Det er to måter å oppsummere søkeresultater på. Forklar hva disse er og hvordan de fungerer.

Svar: Det viktigste her er at studentene kan forskjellene mellom det dynamiske og statiske prinsippet.

Oppgave IV (20%)

Svar rett/galt på følgende utsagn. Hvert **riktig** og **begrunnet** svar belønnes med **2** poeng. **Feil svar** får **-1,5** poeng. **Ubegrunnet** eller **ingen svar** gir **0** poeng.

1. Piksel-til-piksel sammenlikning av to bilder er godt egnet til å beregne/evaluere bildenes likheter.
(~~Rett~~/Galt); Tar ikke hensyn til plassering av piksler. To bilder med samme piksler ikke nødvendigvis like.

2. To lydfiler kan sammenlignes ved å bruke frekvensspektrene til lydene.
   (Rett/~~Galt~~) For eksempel skille mellom tale og musikk…
3. R-frame og R-precision er to mål som brukes i evaluering av søkesystemer.
   (~~Rett~~/Galt) kun R-precision, R-frame er feature i video.
4. Statistisk Thesaurus er et thesaurus som brukes til å forbedre søk eller utvidelse av spørringer når man tar i bruk "*local automatic analysis*".
   (~~Rett~~/Galt) Til global automatic analysis
5. *Harvest* og *Crawler* har samme funksjon, men brukes i to forskjellige web-søkemotorarkitekturer.
   (Rett/~~Galt~~) Harvest til dist. søkemotorer mens Crawler til sentraliserte motorer.
6. Video-informasjon kan ikke gjenfinnes ved hjelp av gjenfinningsmetoder som er laget for bilder og lyd.
   (~~Rett~~/Galt) Kan bruke fargehistogrammet som i bildegjenfinning til å segmentere video til shots
7. Rangering av resultater i vektorbaserte og sannsynlighetsbaserte søkesystemer bruker samme prinsipp.
   (~~Rett~~/Galt) Den en bruker Cos-funksjon som basis, den andre bruker sannsynlighetsberegninger
8. Audiogjenfinningssystemer kan bruke teknikker kjent fra tekstgjenfinngsystemer.
   (Rett/~~Galt~~) Etter talegjenkjenning kan dette brukes.
9. Bevegelsesinformasjon er ganske nyttig som feature til videogjenfinning.
   (Rett/~~Galt~~) Bevegelsesinformasjon kan hentes ut fra video og lagres som forteller litt om innholdet i video
10. Bildehistogram kan brukes til å gjenfinne bilder. I tillegg kan det brukes i videogjenfinning.
    (Rett/~~Galt~~) Se svaret i 6.