

Afleveringsopgave i Kunstig Intelligens og Intelligente Systemer, RUC/Datalogi, efterår 2004

Besvarelsen afleveres i 3 eksemplarer til kursusansvarlig underviser Jørgen Villadsen senest fredag den 19. november 2004 kl. 15 (idet den tidligere officielle frist kl. 12 er under ændring til kl. 15)

Besvarelsen, der gerne må udfærdiges i grupper på op til 3 medlemmer, er en del af den mundtlige eksamen (som i øvrigt er individuel). Navnene på gruppemedlemmerne skal fremgå af forsiden, og siderne i besvarelsen skal være fortløbende nummererede. Besvarelsen skal bl.a. indeholde:

- Diskussion af design, herunder begrundelse for væsentlige valg og afgrænsninger.
- Testkørsler, der overskueligt dokumenterer den implementerede funktionalitet.
- Kortfattet konklusion.
- Kildekoden (en ZIP-fil med kildekoden skal umiddelbart fremsendes på forlangende).

Introduktion

Efter flere møder med firmaet Dorthes Bog Centrum (DBC) er der udarbejdet en samarbejdsaftale og en kravspecifikation, hvori der forligger rammer, ønsker og krav til udvikling af en prototype, der skal undersøge muligheden for avanceret søgning i firmaets database. DBC har p.t. et traditionelt søgesystem, der er outsourcet, hvorfor det ikke er muligt at benytte elementer fra dette i forbindelse med prototypen.

DBC har ca. 77.000 bøger (bibliografiske enheder), som alle ønskes søgbare fra prototypen. Datagrundlaget for prototypen er tabellen DBC med ca. 250.000 poster på formen:

ID	TITEL	AAR	FORFATTER	EMNEORD
157745	Narcissisme	1983	møhl, bo	psykologi-undervisning
157745	Narcissisme	1983	thielst, peter	psykologi-undervisning
157745	Narcissisme	1983	møhl, bo	samarbejdslære
157745	Narcissisme	1983	thielst, peter	samarbejdslære

ID er bogens unikke identifikation, AAR udgivelsesåret og EMNEORD er tilknyttede emneord. Bemærk, at datagrundlaget ikke er normaliseret. For at få adgang til DBC-tabellen og øvrige tabeller og funktioner skal man benytte KIS som præfiks (altså KIS.DBC) på databaseserveren: <http://dat-db.ruc.dk>

DBC forventninger til en kommende prototypes funktionalitet kan opsummeres til følgende:

DBC er træt af Common Command Language (CCL) i det eksisterende system og vil derfor gerne have noget der er mere fleksibelt / intelligent. Man skal således kunne indtaste en beskrivelse, i form af en liste af ord, uden brug af logiske operatoren, og få et svar, der er ordnet med de mest relevante bøger først.

Det lægges ikke vægt på grænsefladen – en simpel linieorienteret dialog er fuldt tilstrækkelig.

Forundersøgelse

I forbindelse med forundersøgelsen til denne prototype har der været foretaget datamining på datagrundlaget for at afklare mulighederne. Resultatet af analysen af emneords ko-forekomster findes i tabellen FM (altså KIS.FM) med poster på formen:

ORD1	ORD2	FM
kv1	foder	44
kvælstof	hovedopgave	44
malkekvæg	hovedopgave	44
metabolism	kv1	44
nitrogen	hovedopgave	44
nitrogen metabolism	kvælstofcyklus	44
svamp	fungi	44

For eksempel optræder “kv1” og “foder” således sammen i 44 forskellige poster.

Der kan med udgangspunkt i denne analyse etableres et mål for association mellem emneord. En associationsfunktion $a(O1,O2)$, der beskriver graden af association fra $O1$ til $O2$, er et mål for sammenfaldshyppighed. Derfor må funktionen først og fremmest afspejle et forhold imellem antallet af sammenfald af $O1$ og $O2$, og antallet af forekomster af $O1$ og $O2$. Som rettesnor for designet af denne funktion skal følgende egenskaber yderligere overholdes:

- Et ord skal associere “fuldt” (til graden 1) til sig selv.
- Intet ord må associere fuldt til et andet.
- “Jo mere sammenfald jo højere grad af association” (hvis A optræder hyppigere sammen med $B1$ end sammen med $B2$, så skal A associere mere til $B1$ end til $B2$).
- Et ord A , der altid er tilknyttet sammen med et andet ord B , bør associere “maksimalt” (til næsten 1) til B .

Associationsfunktionen medfører et såkaldt associationsnet over emneord. Man kan overveje at anvende en minimumsværdi (f.eks. 0,1) for associationsgraden imellem emneord, således at der ikke registreres association ved grader under denne værdi i nettet. Denne afgrænsning kan reducere størrelsen af nettet betragteligt.

Lidt notation: Lad p være en funktion, der for et emneord giver mængden af poster, der har emneordet tilknyttet, og lad $|X|$ være en funktion, der for en mængde X giver antallet af elementer i X . $|p(\text{“tamil”})| = 17$ betyder således, at 17 poster har emneordet “tamil” tilknyttet.

Desuden stilles en PL/SQL-funktion til stamformsbestemmelse af ord til rådighed, jævnfør følgende lille eksempel:

```
SQL> select kis.func.stamform('biler') from dual;
```

```
KIS.FUNC.STAMFORM('BILER')
```

```
-----  
bil
```

Krav til prototypen

Det er i prototypen tilstrækkeligt kun at søge i bøgernes emneord.

I forbindelse med evalueringen af en forespørgsel skal associationsnettet benyttes til ekspansion af forespørgslens termer. Der skal konstrueres en funktion, der kan ekspandere en term til en såkaldt fuzzymængde af associerede termer, hvor medlemsgraden svarer til graden af association. Eksempelvis kunne termen "bil" tænkes at ekspandere til følgende fuzzymængde af associerede termer:

$$\text{ekspand("bil")} = \{1.0/\text{bil}+0.6/\text{bus}+0.3/\text{taxa}\}$$

Der skal således udvikles et søgesystem, der kan evaluere en forespørgsel i form af en mængde af termer. Hertil skal associationsnettet naturligvis udnyttes til at danne svar i form af fuzzymængder, der naturligt kan ordnes efter medlemsgraden. Svaret skal altså være en ordnet liste af bøger, angivet med den medlemsgrad, hvormed de opfylder forespørgslen.

OWA (Ordered Weighted Averaging) skal benyttes som aggregeringsfunktion. OWA er baseret på en vektor af n vægte $W=[w_1, w_2, \dots, w_n]$, hvor vægtene er tal i intervallet $[0,1]$ og summen af vægtene er 1. OWA aggregerer værdierne a_1, a_2, \dots, a_n som summen fra $j=1$ til n af $w_j * b_j$, hvor b_j er den j 'te højeste værdi fra a_1, a_2, \dots, a_n .

Hvis muligt ønskes – udover forespørgsel/svar-delen – funktionalitet i grænsefladen til justering af vægtene i W .

Det forventes, at eventuelle uklarheder i opgaven diskuteres, og at JDBC anvendes.