

Automatisk bestemmelse af formatunderstøttelse

Synopsis

151082 | Søren Vrist (seet@diku.dk)
260576 | Sune Juel Jensen (di050504@diku.dk)

24. Februar 2006

Indhold

| | | |
|---|---------------------|---|
| 1 | Baggrund | 3 |
| 2 | Problemformulering | 4 |
| 3 | Afgrænsning | 5 |
| 4 | Arbejdsopgaver | 6 |
| 5 | Tidsplan | 7 |
| 6 | Rapport-disposition | 9 |

1 Baggrund

Det kongelige bibliotek¹ har igangsat et projekt for at indsamle og bevare det danske internet[Netarkiv] og indsamler i den forbindelse store mængder dokumenter i mange forskellige formater. Her bruges ordet "dokument" i den bredest mulige forstand, f.eks. incl. mediefiler og eksekverbare filer. I den forbindelse taler man om begrebet *formatunderstøttelse* som det at kunne vise et dokument som oprindeligt tiltænkt. Ordet "vise" bruges ligeledes i bredest mulige forstand, f.eks. afspille en mediefil eller vise en rapport².

Automatisk afledning af formatunderstøttelse er dybt essentiel inden for digital arkivering. Automatisk afledning er automatisk at kunne finde en måde, enten direkte at vise et dokument på, eller konvertere og derefter vise dokumentet, via en eller flere mellemlid. Oftest tales der om at beslutte et tradeoff ved enten at altid have en "viewer" til alle formater eller altid konvertere til et fælles basis format. For at få et overblik over problemet vil automatisk afledning af formatunderstøttelse være en stor hjælp. Dette har resulteret i at det kongelige bibliotek har tilbudt at indgå i et samarbejde med DIKU som blandt andet skal resultere i denne bacheloropgave.

Analogt til bootstrapping af oversættere som illustreres med Bratman T-diagrammer, foreslår KB at processen fra et dokument til visning kan illustreres som et T-diagram, hvor de forskellige dele skal passe sammen som dominobrikker. I den forbindelse ønsker KB udviklet en grammatik til at beskrive T-diagrammer med henblik på løsning af en række specifikke opgaver i forhold til netarkivet og formatunderstøttelse.

Fra et datalogisk synspunkt er opgaven interessant fordi der skal anvendes en række kendte datalogiske begreber til løsning af et håndgribeligt problem på virkelig data. Derudover har vi måttet konstatere at der er en udpræget mangel på materiale om implementering og praktisk anvendelse af T-diagrammer.

¹herefter KB

²Her skal dokument forstås så bredt at det eksempelvis godt kan være uoversat C-kode som "vises" ved at oversætte det og afvikle det på en maskine

2 Problemformulering

Kan man konstruere og implementere en syntaks og en eller flere semantikker for T-diagrammer, som kan benyttes i sammenhæng med mere avancerede applikationer hvor T-diagrammer indgår som en essentiel komponent? Og kan man i den sammenhæng udvikle og implementere en eller flere sådanne applikationer, eksempelvis en grafisk baseret applikation til visualisering af diagrammerne ?

3 Afgrænsning

Afgrænsningen er del op i to afdelinger. Første afdeling er udspecificering af rammer for rapport og programmel og anden afdeling er afgrænsning af de specifikke opgaver i form af successkriterier.

Vores mål er primært at skrive simpel og funktionel programkode, vi vil ikke lægge vægt på at skrive optimal eller videre avanceret kode. Al programkode vil udelukkende være kommenteret med henblik på intern forståelse i opgavegruppen, og vi vil ikke udarbejde anden dokumentation til udviklede programmel end kommentarerne og rapporten. I slutningen af hvert af rapportens kapitler dokumenteres et eksempel af en kørsel som demonstrerer at programmet virker og giver et eksempel på hvordan det bruges.

Programteksten offentliggøres på en hjemmeside og vedlægges rapporten som appendiks. Rapporten skrives på dansk.

Vi har valgt Java J2SE som implementationssprog på baggrund af et ønske fra KB og da vi er fortrolige med det på forhånd. Vi designer alle applikationer til afvikling på DIKUs system, da KB ikke har stillet et udviklingsmiljø til rådighed og J2SE sikrer en rimelig grad af portabilitet.

Der udvikles ingen præsentationslogik men kun API'er som faciliterer brugen af udviklet programmel.

Vore succeskriterier for opgaven er følgende:

- Udvikling af en grammatik som kan udtrykke T-diagrammer og deres sammensætning.
- Implementation af grammatikken, i form af et modul som kan parse et diagram og afgøre om det er lovligt. Implementeres i Java.
- At beslutte og beskrive semantikker for T-diagrammer til løsning af de efterfølgende nævnte opgaver. Mindst en af disse semantikker beskrives formelt.
- At implementere en måde at kontrollere et program i vores grammatik er korrekt.
- En grafisk baseret applikation til visualisering af T-diagrammer.

Hvis opgavens tidshorisont tillader det, er det vores ambition at implementere en eller flere af følgende applikationer, efter at vore succeskriterier er opfyldt:

- Definere databaseformat til brug i løsningen af nedenstående opgaver.
- Givet en database af formater udtrykt som T-diagrammer, en beskrivelse af en maskine og et program eller dokument, udskriv alle eksisterende måder at afvikle programmet på, eller vise dokumentet.
- Givet en database af formater udtrykt som T-diagrammer, en beskrivelse af en maskine og et program eller dokument, afgør om afvikling af programmet eller visning af dokumentet er redundant understøttet.
- Evt. andre applikationer, som defineres af KB.

Disse er valgt da de er blandt de problemer KB ønsker løst.

4 Arbejdsopgaver

Arbejdsopgaverne skal ses som en udspecificering af problemformulering og afgrænsning til en række opgaver der skal udføres.

Vi har delt arbejdsopgaverne op i tre overordnede områder.

- **Grammatik**

Grammatikken indebærer at definere et programmeringssprog til at beskrive T-diagrammer og ved implementation at kunne lexe og parse programmer i vores grammatik.

- Planlægning, diskussion, beslutningstagen, som munder ud i eksempler med tilhørende parsetræer og en generel BNF.
- Implementation. Dvs. en lexer og en parser
- Rapport

- **Semantik**

Semantik indebærer de forskellige udtryksmuligheder grammatikken kan lede til.

- Planlægning, definition, diskussion, beslutningstagen omkring semantik generelt.
- Kontrol af at parsetræ opfylder domino-egenskab for T-diagrammer implementeres, en kontrolenhed.
- En "tegner" implementeres oven på grammatikken.
- Rapport

- **Query-engine**

Query-enginen er den brede overskrift til at kunne få svar på spørgsmålet om "givet en server og et program, på hvor mange måder kan jeg få afviklet programmet" og "givet en server, hvilke formater er der ikke redundant mulighed for at få afviklet". Dette kræver endnu en semantik til vores grammatik.

- Planlægning, diskussion, beslutningstagen
- Implementation
- Rapport

- **Resten af rapporten**

- Indledning
- Konklusion
- Related work
- Litteraturliste

5 Tidsplan

Det er forsøgt at lægge arbejdspresset mest muligt på den. sidste af de to studieblokke, da vores tid er mest begrænset af andre faktorer i den første blok.

- **Februar**

2006-02-24 Aflevere synopsis

- **Marts**

2006-03-01 Forsvare synopsis

2006-03-24 Milepæl nr 1.

- * Første udkast til grammatik, i form af eksempler, parsetræer og BNF.
- * Valgt lex og yacc.
- * Påbegyndt semantikovervejelser til kontrolenhed.

- **April**

2006-04-28 Milepæl nr 2.

- * Evt. nyt grammatikudkast
- * Første implementation af grammatikken i form af lexer,parser
- * Første udgave af kontrolenhed.
- * Påbegyndt "tegner"-semantik.
- * Påbegyndt "tegner"-implementation
- * Første rapportafsnit om grammatik og implementation af denne

- **Maj**

2006-05-19 Milepæl nr 3.

- * Evt. grammatik tilpasning
- * Evt. parser tilpasning
- * Evt. tilpasning af rapport om grammatik og parser.
- * Første implementation af "tegner"-semantik.
- * Afsnit om "tegner" og kontrolenhed-semantik i rapporten
- * Første udkast til query-engine-semantik.
- * Påbegyndt impl. af query-engine.

- **Juni**

2006-06-05 Milepæl nr 4. Code-freeze.

- * Grammatik færdig, implementeret og beskrevet i rapport
- * Kontrolenhed beskrevet og afgrænset i rapporten
- * "tegner" færdig, implementeret og beskrevet i rapport
- * Query-engine færdig, implementeret og rapportafsnit påbegyndt

2006-06-16 Rapport færdig...korrekturlæsning, Milepæl 5
Rapportafsnit ud over allerede færdige afsnit

2006-06-19 Korrekturlæst, aflevere rapport

2006-06-28 Forsvare rapport

Tidsplanen skal ikke ses som en krakilsk sekventiel forløb men nærmere et iterativt forløb hvor vi løbende arbejder på en eller flere opgaver som symboliseret ved denne matrix:

| | M1 | M2 | M3 | M4 | M5 |
|-------------------|----|----|----|----|----|
| Grammatik | 1 | 2 | 3 | 4 | |
| Grammatik-impl | A | 1 | 2 | 3 | |
| Grammatik-rap | | 1 | 2 | 3 | 4 |
| Kontrolenhed | A | 1 | 2 | 3 | |
| Kontrolenhed-impl | | 1 | 2 | 3 | |
| Kontrolenhed-rap | | | 1 | 2 | 3 |
| Tegner | | A | 1 | 2 | |
| Tegner-impl | | A | 1 | 2 | |
| Tegner-rap | | | 1 | 2 | 3 |
| Query | | | 1 | 2 | |
| Query-impl | | | A | 1 | |
| Query-rap | | | | | 1 |
| Code-Freeze | | | | X | X |
| Misc. rapport | | | | | 1 |

A betyder påbegyndt, og tallene repræsenterer udgave/iteration af den pågældende opgave.

6 Rapport-disposition

- Indledning
- Konklusion
- Grammatikken
- Et afsnit for hver semantik
- Kort systembeskrivelse
- Related work
- Litteraturliste
- Appendix(Kildekode)

Appendix

Litteratur

[Netarkiv] Netarkivet, <http://netarkivet.dk/index-da.php>

[Opgave] Opgavebeskrivelsen, <http://www.diku.dk/simon-sen/bach/KBDIKU05diagrams/> simon-

99