

Synopsis

Projekttitel: Automatisk (grov)sortering af dokumenter og websider efter indhold

Af Morten Gryning 110484 og Mikkel Mertz 090883

Problemformulering

Kan man med et softwareprogram til dokumentklassificering kategorisere dokumenter efter indholdet af fortroligt materiale og med hvilken nøjagtighed?

Med nøjagtighed menes i hvor stor grad programmet kategoriserer dokumenterne i fortroligt og ikke fortroligt materiale korrekt.

Afgrænsning

Programmet skal fungere ved slutningen af projektet, men behøver ikke kunne anvendes uden for testmiljøet.

Der behøver ikke at blive skrevet nogen brugerhjælp til programmet.

Der behøver ikke udvikles et helt nyt program – det eksisterende program kan udvides og undersøges.

Nøjagtighed for programmets kategorisering behøver ikke bedømmes af andre end Det Kongelige Bibliotek. Nøjagtighedskriterier defineres af Det Kongelige Bibliotek.

Det er ikke en målsætning at opnå større viden indenfor de forskellige kategoriseringsmetoder, men blot benytte den eller de, der findes bedst egnet at implementere og benytte i opgaven.

Begrundelse

Det Kongelige Bibliotek indsamler og bevarer store samlinger af dokumenter og websider. Under denne indsamling indsamles bl.a. fortroligt og ikke offentligt tilgængeligt materiale under henhold til pligtaflevering, der ikke må tilgås af andre end forskere under tavshedserklæring. Den manuelle del af frigivelsesarbejdet kan lettes betragteligt ved at lade et program lave en (grov)sortering af materialet først.

Evaluerings

Der skal foretages test af programmet med en række pseudodokumenter eller rigtigt fortroligt materiale. Dette skal teste nøjagtigheden af programmets kategorisering. Evalueringen af testen sker ved sammenligning mellem programmets kategorisering og Det Kongelige Biblioteks kategorisering. Sammen med Det Kongelige Bibliotek evalueres om programmets nøjagtighed er tilfredsstillende. Det Kongelige Bibliotek kan være tilfreds med en grov kategorisering, der giver fejl, men som viser et potentiale for prototypens koncept.

Arbejdsopgaver

Analyse af det eksisterende program og tilgængelige kategoriseringsmetoder

Sammen med Det Kongelige Bibliotek undersøges deres eksisterende program og tilgængelige kategoriseringsmetoder undersøges ud fra litteratur på området i samspil med Det Kongelige Bibliotek. Dette skal føre til en beslutning om hvorvidt et nyt program konstrueres eller det eksisterende udvides. Derudover vælges også en eller flere brugte kategoriseringsmetoder i programmet ud fra betragtninger om deres styrker og svagheder, samt krav til udvikling. Analysen skal kun bruges til at vælge og begrunde valget af en eller flere kategoriseringsmetoder, og ikke være en detaljeret analyse om metoderne.

Fremstilling af prototype eller udvidelse af det eksisterende program

Efter analyse af kategoriseringsmetoder og det eksisterende program er færdigt, fremstilles en prototype med de valgte kategoriseringsmetoder.

Test af det implementerede programs nøjagtighed

Efter at programmet til kategorisering er udviklet, testes dets nøjagtighed i samarbejde med Det Kongelige Bibliotek. Der vil blive benyttet enten pseudodokumenter eller rigtige fortrolige dokumenter til denne test.

Skrivning af rapport

Efter at programmet er blevet testet, skrives rapport over udviklingen af programmet og dets nøjagtighed.

Metoder

Test af det implementerede programs nøjagtighed

Til testen af nøjagtighed samarbejdes med Det Kongelige Bibliotek for at bedømme programmets nøjagtighed og evt. justeres programmets kategorisering ud fra denne test.

Fremstilling af prototype eller udvidelse af et eksisterende program

Til udvikling af programmet benyttes programmeringssproget JAVA og udviklingsværktøjet Eclipse.

Skrivning af rapport

Der vil blive skrevet rapport på dansk. Kildekodens kommentater vil enten blive skrevet på dansk eller engelsk – det vælges sammen med Det Kongelige Bibliotek.

Information og informationskilder

For at få information omkring eksisterende kategoriseringsmetoder læses bl.a:

- http://en.wikipedia.org/wiki/Machine_learning.

- ”Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging, Department of Computer Science”, 1995, Johns Hopkins University, Eric Brill”.

For at få information om det eksisterende program og læses bl.a.:

- ”File Format Identification and Characterization by Rule Inference”, dato ukendt, Anders S. Johansen, Royal Library, Denmark.

Tidsplan

23. februar: Første milepæl. Fastsættelse af hoved- og delmål for projektet.

28. februar: Synopsforsvar på DIKU.

2. og 9. marts: Ugentlige møder med Det Kongelige Bibliotek om fredagen af op til to timers varighed hver på DIKU.

16. marts: Anden milepæl. Afsluttet arbejdsopgave ”Analyse af det eksisterende program og tilgængelige kategoriseringsmetoder”.

30. marts, 6. og 13. april: Ugentlige møder med Det Kongelige Bibliotek om fredagen af op til to timers varighed hver på DIKU.

20. april: Tredje milepæl. Midtvejsevaluering. Projektet evalueres og det besluttet om projektet skal udvides, beskæres eller forsættes uændret.

27. april, 4. og 11. maj: Ugentlige møder med Det Kongelige Bibliotek om fredagen af op til to timers varighed hver på DIKU.

18. maj: Fjerde milepæl. Koden fryses og kun nødvendige ændringer foretages. Projektet evalueres igen og omfang af projektet kan blive justeret.

25. maj, 1. og 8. juni: Ugentlige møder med Det Kongelige Bibliotek om fredagen af op til to timers varighed hver på DIKU.

13. juni: Rapportaflevering.

13. juli: Første generalprøve på fremlæggelse.

20. juli: Anden generalprøve på fremlæggelse.

27. eller 28. juli: Projektfrelæggelse.

Risikovurdering

Hvis analyse arbejdsopgaven går galt, kan det risikeres at der ikke findes nogen brugbare og eksisterende kategoriseringsmetoder, der vil give tilstrækkelig nøjagtighed ved kategorisering af dokumenter. Dette vil betyde at egen kategoriseringsmetode må forsøges udviklet og benyttet. Dette kan medføre dårligere nøjagtighed end krævet af Det Kongelige Bibliotek.

Hvis det viser sig at implementering af en valgt kategoriseringsmetode tager for lang tid, kan en mere simpel kategoriseringsmetode benyttes. Der vil vælges at få udviklet et simpelt program først, der kan udvides senere i projektet.

Hvis evalueringen af prototypens nøjagtighed i kategoriseringen af dokumenter og websider ikke er tilstrækkelig for Det Kongelige Bibliotek, kan systemet forsøges at blive ændret. Hvis det viser sig at systemet ikke kan blive ændret på det nuværende tidspunkt vil der redegøres for hvorfor den

ønskede nøjagtighed ikke kunne opnås.

Disposition for rapporten

Det vil forsøges at opnå følgende fordeling af sider i rapporten:

- Baggrund (2 sider)
- Problemformulering og afgrænsning (2 sider)
- Analyse (10 sider)
- Design af system (12 sider)
- Implementering (5 sider)
- Test (8 sider)
- Diskussion og konklusion (8 sider)
- Litteratur (1 sider)
- Bilag (ikke angivet sideantal)
 - Testresultater
 - Kildekode
 - Andet

Dette giver i alt 48 sider uden bilag.