

Prosjektbeskrivelse

“The goal of this project is to find ways to optimize the crawler frequency for individual web sites. The idea is to avoid crawling a site in case the accessibility to the site has not been changed. This is a challenge for all search engines to focus the resources on actual changes. Another relevant aspect of sampling is to select a significant and representative set of sites.”

Etter samtale med veilederne har vi kommet fram til å konsentrere oss om forandringer i tilgjengelighet. Dermed kan vi utelukke endringer i tekstlig og grafisk innhold. Vår oppgave er å hente ut taggene fra HTML-dokumentet og deretter sette en karakter på i hvor stor grad disse følger W3Cs retningslinjer for tilgjengelighet.

Konkret innebærer dette at gruppemedlemmene må sette seg inn i programmeringsspråket Python. Dessuten kommer vi til å jobbe mot en database (MySQL eller PostgreSQL) hvor både de innsamlede og beregnede dataene blir liggende. For å sette karakteren, har vi benyttet oss av W3C sine retningslinjer beskrevet i dokumentet “Web Content Accessibility Guidelines” [<http://www.w3.org/TR/WCAG20/>]. Der er det skissert 14 retningslinjer med tre nivåer som beskriver tilgjengeligheten til en bestemt webside.

Vi har sett for oss en parser som grupperer sidene i forhold til oppdateringsfrekvensen som blir registrert. Vi kommer til å definere en startverdi for denne frekvensen, og dersom en bestemt side er oppdatert neste gang parseren kommer innom, vil tidsintervallet halveres. I motsatt fall vil tidsintervallet fordobles. På lengre sikt vil det kanskje være interessant å finjustere denne algoritmen til også å ha mulighet for andre intervaller.

Som sammenligningsgrunnlag for tilgjengelighet har vi valgt den webbaserte klassifiserings-tjenesten Bobby [<http://bobby.watchfire.com/>].

Arbeidsplan

Vi ser for oss fire faser under produksjonen av crawleren:

- 1) Innsamling av HTML-dokumenter
- 2) Hashverdi beregnes – grovtesting for å skille ut statisk innhold
- 3) Grovrensing for å rense bort alt utenom taggene
- 4) Analysing av taggene i forhold til kriteriene i WCAG, resultatene beregnes og legges i databasen

Vi har ikke ennå tatt stilling til hvem som får hvilke oppgaver i de avsluttende fasene. Foreløpig fremdriftsplan ser slik ut:

Fase	Carl T. Vatne	Lars R. Haugen	Per Ø. Hodøl	Ferdig innen
Forberede 30.09	Samarbeid	Eksamen	Samarbeid	30.09.2004
Sette seg inn i Python	Samarbeid	Samarbeid	Samarbeid. Lånt to bøker	05.10.2004
Innsamling av HTML-data		Skaffe liste over aktuelle URL'er	Programmere modul	05.10.2004
Lage db-tabellene	Gjør det som trengs for db-tilgang i oppgaven			05.10.2004
Hash-testen		Programmere modul		12.10.2004
Parse HTML-data og returnere tagger	Samarbeid			19.10.2004
Analysere tag-innhold, gi karakter	Samarbeid			02.11.2004
Legge karakterene i database-tabell	Samarbeid			02.11.2004
Lage dokumentasjon over systemet	1/3 av systemet	1/3 av systemet	1/3 av systemet	16.11.2004
Skrive prosjektrapporten	Samarbeid			02.11-16.11
Forberede presentasjonen	Samarbeid			16.11-25.11