Projektansøgning til Center for IT-forskning

Projekt titel:

Anvendt web usage mining af computer mediated communication

Deltagere:
Institut for Kommunikation, Journalistik og Datalogi, RUC
E-sense A/S
Virksomhed X

Sammenfatning:

Projektet søger om et samlet beløb på kr. 1.341.200,- fra CIT. Det skal understøtte projektets budget på kr. 3.016200,-, hvor E-sense bidrager med kr. 1.275.000,-, Institut for Datalogi, RUC bidrager med 200.000,- og Virksomhed X bidrager med kr. 280.000,-. Virksomhed X er ikke identificeret på ansøgningstidspunktet, men konkrete forhandlinger pågår med Teledanmark A/S og Nesa A/S.

Projektet skal bidrage til forskningen i og udnyttelsen af adfærdsdata fra Internettet. Bidraget skal dels være i form af fortolkningsrammer, metoder og teknikker, som har relevans for forskningen, dels udvikling og tilpasning af praktiske metoder og heuristikker, som kan udbrede udnyttelsen af adfærdsdata i samfund og erhvervsliv.

De videnskabelige mål med projektet er:

1. at kortlægge og evaluere eksisterende metoder og produkter til web usage mining
2. at eksperimentere med udvikling af nye analysetyper til web usage mining med fokus på Computer Mediated Communication
3. at udvikle og forbedre metoder til anvendelse af web usage mining i praksis og integrere det med andre datakilder.

De kommercielle mål med projektet er:

1. At øge viden om muligheder og begrænsninger ved web usage mining.
2. At implementere teknikker og metoder til web usage mining.
3. At udvikle servicekoncepter og konsulentydelser, som udnytter web usage mining teknikker og metoder.

De kommercielle resultater af projektet vil bestå af metoder, servicekoncepter og konsulentydelser. Det er ikke målet med projektet at udvikle egentlige produkter.

Baggrund:

Digitaliseringen af kommunikation generelt og forretningstransaktioner specifikt giver en eksplosion af muligheder for at analysere kommunikation og forretningstransaktioner og anvende analyserne på forskellig måde.
Det har betydet skabelsen af disciplinen Data Mining (Hand, Mannila et al. 2001). Data Mining disciplinen beskæftiger sig generelt med udvikling af algoritmer og praktisk gennemførsel af analyser af forskellige sammenhænge i store datamængder. Anvendelse af Data Mining teknikker til anvendelse i drift og udvikling af en forretning kaldes generelt Business Intelligence.
Opkomsten af WWW har betydet en eksplosion i mulighederne for at analysere data. Det har betydet opkomsten af en disciplin under Data Mining, som beskæftiger sig med mining af data fra WWW(Cooley, Srivastava et al. 1997; Kosala and Blockeel 2000). Den disciplin kan overordnet opdeles i *web content mining*, som analyserer indholdet på WWW og *web usage mining* som analyserer log-data om anvendelsen af WWW. Nærværende projekt beskæftiger sig med *web usage mining*.
Web usage mining teknikker bliver i dag anvendt på forskellige områder, men potentialet er langt fra udtømt.

Web usage mining disciplinen er udsprunget af data mining og database discipliner. Det er traditionelt naturvidenskabeligt funderede discipliner, som fokuserer på løsning af veldefinerede algoritmiske problematikker. Analyser af web anvendelser kræver dog en høj grad af tolkning og overvejelser omkring metodisk integration med andre undersøgelses-former. Det kan både være surveys, fokusgrupper, dokumentanalyse og interviews. Det er et fokuspunkt for dette projekt at forsøge at gøre generelle analyseteknikker som sekvensanalyser, clustering etc. - der er vel beskrevede som Data Mining teknikker - anvendelige i praktiske analyser af web-baseret kommunikation. HCI (Human-Computer-Interaction) er det felt indenfor forskning, som hidtil i nogen grad har beskæftiget sig med anvendelsen af web usage mining teknikker.
HCI's anvendelse af data mining er fokuseret omkring anvendelsen af log-analyser som input til design af den enkelte brugers interaktion med en web-site. Den mest udbredte metode er click-stream analyser, som er baseret på sekvens-analyser. Click-stream analyser anvendes bl.a. til overvejelser omkring re-design af web-siten (Spiliopoulou, Pohle et al. 1999; Mobasher, Cooley et al. 2000).

I (Bøving Forthcoming 2002) eksperimenteres med metoder til analyse af Computer Mediated Communication (CMC), hvor http-logs anvendes til identifikation af genrer af kommunikation. Indenfor rammerne af DIWA forskningsprogrammet (www.diwa.dk) har to af projektansøgerne, Jesper Simonsen og Kristian Billeskov Bøving arbejdet med analyser af log-filer som en del af et brugsstudie af en web-baseret samarbejdsapplikation. Disse analyser viser en ny anvendelse af log-filer til studier af kommunikation mellem sociale aktører.
Analyserne kan udnyttes som input til at studere CMC som et socialt fænomen og dermed bidrage til forskningen i de sociale strukturer som opstår på WWW.

Resultaterne af en CMC log-analyse kan også finde anvendelse i mere praktisk orienterede sammenhænge. De kan anvendes på meget forskellige måder i en web-sites livscyklus:

- som input til design- eller redesign-proces
- til evaluering af målsætninger
- som input til efterfølgende brugersession (personaliseringer)
- Som input i den samme brugersession (dynamisk tilpasning af bruger-interfacet)

Formål

Projektet har både videnskabelige og kommercielle formål.
Det videnskabelige formål med projektet er:

1. at kortlægge og evaluere eksisterende metoder og produkter til web usage mining
2. at eksperimentere med udvikling af nye analysetyper til web usage mining med fokus på Computer Mediated Communication
3. at udvikle og forbedre metoder til anvendelse af web usage mining i praksis og integrere det med andre data.

Ad 1. Web usage mining findes dels beskrevet i forsknings-litteraturen(Cooley, Srivastava et al. 1997; Kosala and Blockeel 2000), dels er den implementeret i en

række software-produkter (f.eks. WebTrends, Clementine fra SPSS), freeware (f.eks. Analog, WwwStat, Weblog, WUM) og service-koncepter (f.eks. redsheriff.com, SurfAid fra IBM). Det giver et uoverskueligt landskab af teknikker og metoder, som ikke er beskrevet systematisk. Det er projektets formål at skabe et kritisk overblik over feltet. Det skal skabe baggrund for udviklingen af fortolkningsrammer for web usage mining. Fortolkningsrammer giver mulighed for at vurdere både validitet og pålidelighed ved forskellige praktiske anvendelser af en analysetype. Det er dels et mål i sig selv, dels et udgangspunkt for at fokusere indsatsen i 2 og 3.
Ad 2. Anvendelse af web usage mining bør ikke kun ske til analyse af enkeltbruger sessioner. Det kan også anvendes til at undersøge WWW som kommunikationsmedium(Bøving Forthcoming 2002). Projektet vil etablere et analysemiljø, som skal eksperimentere med udviklingen af nye analysetyper.
Ad 3. Det er et væsentligt kriterium for analysetyper, at de integreres i livscyklus'en på den web-site, som de udføres på. Det kan være til personalisering, recommender systems, re-design eller til evaluering af målsætninger. Ved udvikling af analyser til undersøgelse af Computer Mediated Communication er det desuden væsentligt at integrere log-data med andre data-kilder som surveys og fokusgrupper og interviews. Udviklingen af metoder vil derfor også være rettet mod sociologiske studier af medieanvendelse.

De kommercielle mål med projektet er:

1. At øge viden om muligheder og begrænsninger ved web usage mining.
2. At implementere teknikker og metoder til web usage mining.
3. At udvikle servicekoncepter og konsulentydelser, som udnytter web usage mining teknikker og metoder.

Ad. 1.
De deltagende virksomheder skal gennem samarbejde i projektet og formidling af resultater øge deres viden om muligheder ved web usage mining og om teknikker og metoder til gennemførsel af konkrete projekter.

Ad. 2.
Gennem projektsamarbejdet vil Virksomhed X implementere udvalgte teknikker og metoder som en del af forretningen.

Ad. 3.
Gennem projektsamarbejdet vil E-sense A/S udvikle og sælge servicekoncepter og konsulentydelser til erhvervslivet.

Projektplan

Projektet inddeles i fire faser, som hver vil løbe over 6 måneder. Efter hver fase mødes styregruppen. Projektet producerer en statusrapport, som evalueres af styregruppen. Evalueringen danner baggrund for opsætning af konkrete mål for næste fase. Det gælder både mål for forskningsprocessen og mål for formidling af projektets aktiviteter og resultater.
Projektet indleverer desuden en status over projektets økonomi til styregruppen.
Projektet vil ad hoc afholde planlægnings- og evalueringsmøder og arrangementer til formidling af resultater.

I projektets budget er inkluderet deltagelse i videnskabelige konferencer, hvor det er målet at formidle projektets resultater. Det vil både være mere specifikke konferencer om data mining (f.eks. SIGKDD) og bredere konferencer om anvendelse af WWW og IT generelt (f.eks. ICIS, A(o)IR)

Fase 1	Mål: Opstart af analysecenter og kortlægning af eksisterende web usage mining metoder. Opgaver: - Ansættelse af assistent - Etablering af analysecenter - Etablering af web-site til løbende formidling af resultater. - Etablering af kontakt til internationale forskningsmiljøer. - Evt. Indgåelse af sponsoraftaler med softwareleverandører - Kortlægning og evaluering af metoder til web usage mining - Formidling af kortlægningen, dels til forskningsmiljøet, dels bredere.
Fase 2	Mål: Valg af analyseobjekt og eksperimenter med analyser af Computer Mediated Communication + implementering af metoder til anvendelse. Opgaver: - Valg af konkrete cases til analyse - Indsamling af data til analyse - Planlægning og gennemførelse af eksperimenter - Udarbejdelse og implementering af anvendelsesorienterede metoder til udnyttelse af eksisterende mining teknikker. - Formidling af anvendelsesorienterede metoder.
Fase 3	Mål: Afslutning af eksperimenter + implementering og evaluering af metoder til anvendelse. Opgaver: - Evaluering af eksperimenter hos relaterede internationale forskningsmiljøer - Implementering og evaluering af metoder til udnyttelse af eksisterende mining teknikker.
Fase 4	Mål: Samling af resultater fra projektet og formidling af forsknings-mæssige resultater i artikler. Opgaver: - Skrivning af artikler - Deltagelse på konferencer

Projektorganisering

Projektet består af 4 parter:

- Institut for Datalogi Roskilde Universitets Center
- E-sense A/S
- Virksomhed X

- Center for IT-forskning

Den praktiske gennemførsel af projektet vil blive drevet af Jesper Simonsen, lektor på Institut for Datalogi, RUC, Kristian Billeskov Bøving som ansættes som post.doc. på Institut for Datalogi, RUC og af en halvtids-assistent, som vil være ansat hos E-sense.
Den tid som er afsat i budgettet til involvering fra E-sense udover halvtids-assisteneten og involvering fra virksomhed X vil blive anvendt til deltagelse i eksperimenter og evaluering af resuktater fra eksperimenterne.
Projektet etablerer et analyse-center i faciliteter stillet til rådighed af E-sense A/S, som anvendes til eksperimenter med nye analysetyper. E-sense A/S og Virksomhed X stiller data til rådighed for analyserne og deltager i gennemførsel af eksperimenter. Integreringen af analysetyperne vil blive afprøvet i praksis i et samarbejde mellem Kristian Billeskov Bøving, E-sense A/S og Virksomhed X.
Projektet vil derfor dels blive gennemført på Datalogi, RUC dels hos E-sense A/S.
Projektet vil blive ledet i det daglige af Jesper Simonsen, RUC og Kristian Billeskov Bøving, RUC.
Til overordnet styring og løbende evaluering af projektet foreslås nedsat en styregruppe bestående af:
Jesper Simonsen, RUC
David Junge, E-sense A/S
Klaus Bruhn Jensen, Københavns Universitet
X, CIT

Formidling af resultater

Formidlingen af projektets resultater vil foregå løbende i projektets levetid.

- Publicering af artikler på internationale konferencer og i internationale peer-reviewede tidsskrifter (2 – 3 artikler)
- Etablering af web-site til løbende publicering af projektets resultater
- Bredere formidling af resultater ved foredrag, artikler etc.

Projektet vil løbende fastlægge mere præcise mål for formidlingen af resultater, som følges op og evalueres af styregruppen.

Projektbudget

	Enhedspris	Enheder	Beløb	CIT	RUC	E-sense	X
Kontorfaciliteter	10000	24	240000			240000	
Løn post. doc.	36626	24	879024	879024			
FIK	879024	0,23	202176	202176			
Løn assistent	500	1670	835000			835000	
Arbejdsinvolvering E-sense	500	200	100000			100000	
Lektorinvolvering RUC	500000	0,4	200000		200000		
Arbejdsinvolvering Nesa	400	200	80000				80000
Konferencer	30000	2	60000	60000			
Hardware	100000	2	200000				200000
Software	100000	1	100000	100000			
Etablering af center + styregruppe	100000	2	200000	100000		100000	
Total			**3096200**	1341200	200000	1275000	280000

Ansøgere

Institut for Kommunikation, Journalistik og Datalogi, Roskilde Universitetscenter
Kontaktperson: Lektor Jesper Simonsen, e-mail: simonsen@ruc.dk URL: http://www.dat.ruc.dk/~simonsen/ eller Kristian Billeskov Bøving, e-mail Kristian@billeskov.dk URL: www.billeskov.dk.

E-sense A/S
Livjægergade 17
2100 København Ø
URL: www.e-sense.dk
Kontaktperson: Adm. Dir. David Junge, tlf. 35441700, e-mail: dj@e-sense.dk.
E-sense er en Internet konsulent virksomhed, som er stiftet primo 2000. Virksomheden løser opgaver indenfor bl.a. strategi, markedsføring og kommunikation via Internettet. Virksomheden har pt. 40 fuldtidsansatte.

X
Den sidste ansøger er ikke endeligt på plads, men der pågår konkrete forhandlinger med Teledanmark A/S og Nesa A/S om deltagelse i projektet, på de betingelser, som er opstillet i projektets budget.

Samarbejdspartnere

Det er projektets plan at opbygge samarbejder med forskningsmiljøer internationalt. De er ikke på ansøgningstidspunktet endeligt på plads.

Litteratur:

Bøving, K. B. (Forthcoming 2002). Mine the gap - a multi-method investigation of web-based groupware use. Institute for Film & Media Studies. Copenhagen, University of Copenhagen**:** 180.

Cooley, R., J. Srivastava, et al. (1997). Web mining: Information and pattern discovery on the world wide web. 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, California.

Hand, D. J., H. Mannila, et al. (2001). Principles of data mining. Cambridge, Mass., MIT Press.

Kosala, R. and H. Blockeel (2000). "Web Mining Research: A Survey." SIG KDD Explorations **2**: 1-15.

Mobasher, B., R. Cooley, et al. (2000). "Automatic personalization based on Web usage mining - Web usage mining can help improve the scalability, accuracy, and flexibility of recommender systems." Communications of the Acm **43**(8): 142-151.

Spiliopoulou, M., C. Pohle, et al. (1999). Improving the Effectiveness of a Web Site with Web Usage Mining. Workshop on Web Usage Analysis and User Profiling (WebKKD99), San Diego, August 1999.