Spamfiltrering basert på statistiske metoder

Trond Reitan og Cecilie D. Widsteen

- Beskrivelse av spam-problematikken
- Grunnen til statistiske løsninger på problemet
- Naiv Bayesiansk filtering
- Kji-kvadrat-metoden
- LSA
- Statistikk og software
- Det spillteoretiske aspektet
- Spammernes mottrekk
- Raffineringer av metodene
- Oppsummering/kommentarer

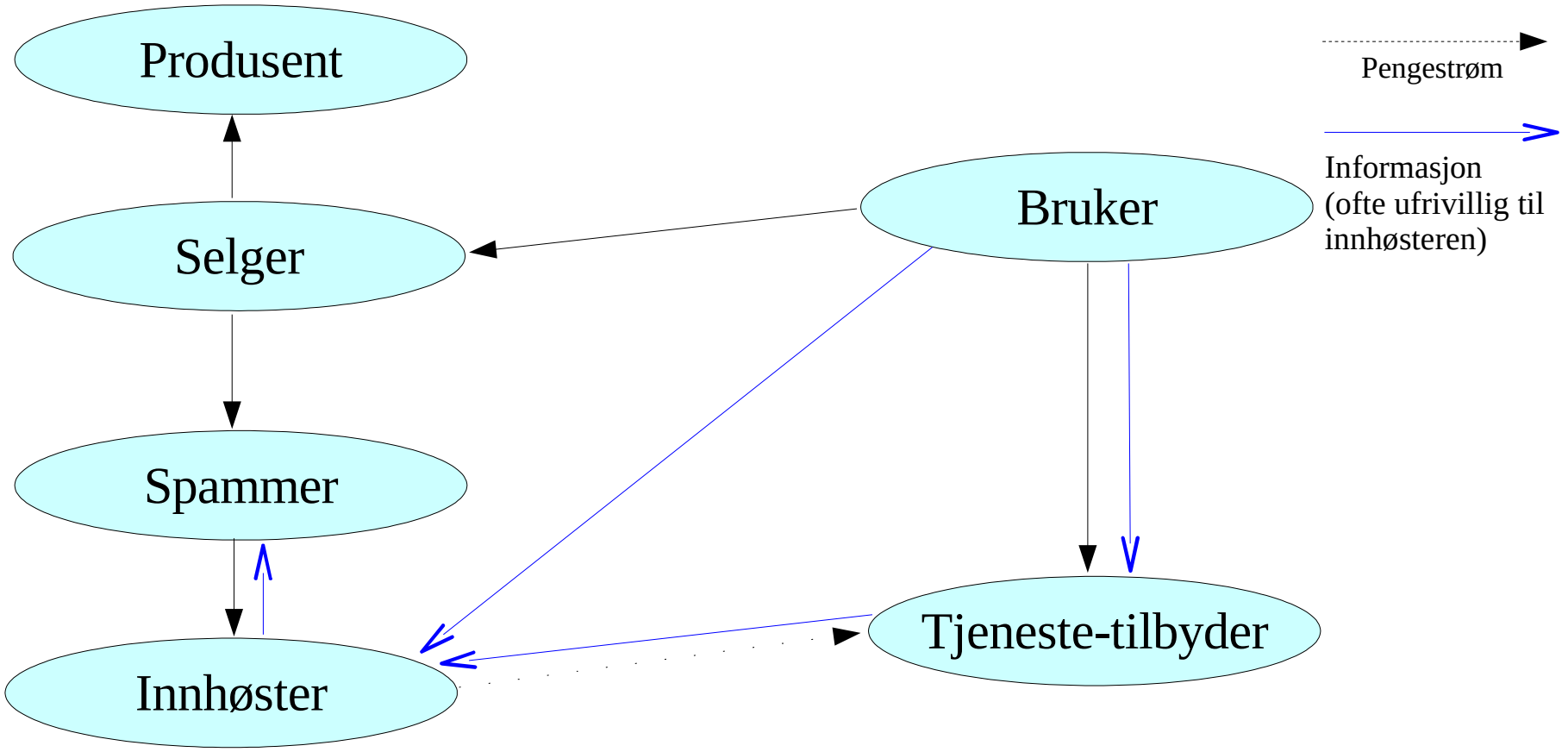
Spammens natur I

- Hva er spam?
  «Masse-utsendelse av epost i den hensikt å promotere/selge et budskap/produkt til mottakerne».
- Spam er uønsket epost.
  - det tar opp lagringsplass
  - har ofte støtende innhold
  - man bruker tid og energi på å skille spam fra vanlig epost, i denne sammenhengen kalt *ham*.
- Derfor trenger man *spam-filtre* Et spamfilter er en programvare som automatisk sorterer ut spam fra vanlig epost.

Spammens natur II

- Forskjellige typer spam-filtre:
  - *Regelbaserte* (heuristiske) filtre; Leter etter mønstre som indikerer spam. Har regler hvor disse mønstrene ikke er tillatt.
  - «*Statistiske*» filtre; Samlebetegnelse for filtre som klassifiserer epost vhja. metoder som sammenligner ordforekomster i innkommende epost med allerede klassifisert epost, gjerne hos brukeren. Tillater brukeren å reklassifisere.
- Statistiske filtre:
  - (Naive) Bayesianske, kji-kvadrat, LSA. I kommersiell bruk fra 2003, stadig mer populært.

Spammens natur - 3
økonomisk struktur



PS: Maks. ekspandert. Ofte kan selger og spammer være den samme. Produktet kan være falske, altså ingen produsent. Der det er en produsent, kan gevinsten fra salg til spam-relaterte selgere være mindre enn tap pga minsket anseelse.
Prikket linje mellom innhøstere og tjeneste-tilbydere kan være (i den grad det gjøres) være salg av epostlister hos internett-tilbydere fra illojale ansatte eller skjult firma-policy.

Hvorfor statistiske løsninger – regelbaserte filtre

- Ulempe med regelbaserte filtre:
  - høy grad av «falske negative», dvs. spam som slipper igjennom.
  - reglene er statiske, må oppdateres etterhvert som «spammere» utvikler nye mønster.
  - spammere kan teste spam på kjente regelbaserte filtre for å komme igjennom.

Hvorfor statistiske løsninger - Praktiske betraktninger

- Praktiske fordeler:

(a) Høyere presisjon

(b)Adaptiv – Filtreringen vil tilpasse seg innholdet i ny spam.

(c)Personlig – Ulike personer mottar ulik spam og har ulike toleransegrenser.

(d)Mottrekk – Siden filteringsgrunnlaget er personlig er det vanskelig for spammere å få tilpasse seg den nye filtreringen.

- Praktiske ulemper: Kan være dårlig i starten. Kan være nødvendig å se igjennom filtrert epost en gang iblant (blir aldri perfekt). Kan fungere som språk-filter.

Naiv Bayesiansk filtrering

- Popularisert av Paul Graham (A Plan For Spam - 2002).
- Metoden ble raskt tatt i bruk. (Eks: Mozillas epost-klient, senere Thunderbird.)
- Man ønsker å finne Pr(epost er spam | epostens innhold).
- Antar a' priori Pr(spam)=Pr(ham)=1/2, og at ordforekomstene er uavhengig av hverandre.
- Har en optimalisert algoritme for å estimere Pr (spam | epost inneholder ord w_i).

Naiv Bayesiansk filtering - 2

- Finner de femten ordene der Pr(spam | w_i) er lengst unna 1/2.
- Bruker Bayes formel begge veier;
- $$Pr(spam|w_{1,}\ldots,w_{15})=\frac{\prod_{i=1}^{15} Pr(spam|w_i)}{\prod_{i=1}^{15} Pr(spam|w_i)+\prod_{i=1}^{15} Pr(ham|w_i)}$$
- Filtrerer når denne sannsynligheten går over en gitt grense (90%). Merk at tapet ved feilaktig spamfiltrering ikke er lik tapet ved feilaktig å slippe spam igjennom.
- Paul Graham rapporterer om 0.03% falske positive og 0.5% falske negative.

Naiv Bayesiansk filtrering - ulemper

- Håndterer ikke ord-avhengighet. Her er det foreslått Bayesianske nettverk som en mulig løsning. (Se også CRM114).
- A priori sannsynlighet for spam, samt andre meta-parametre, bør kunne bestemmes fra brukerens tidligere epost-korpus. Kan håndteres som hierarkiske parametre?

Kji-kvadrat-metoden

- Baserer seg på de samme anslagene av Pr (spam | w_i), men benytter de som p-verdier for nullhypotesen om at en epost er spam.
- For en samling av m tester;
- $p_{total} = C_{2n}^{-1}(-2 \sum_{i=1}^{m} \ln(P_i))$
- Kan danne seg p-verdi både for null-hypotesen «epost er spam»(S) og «epost er ikke spam»(H).
- Kan filtrere på I=(1+H-S)/2, med I rundt ½ som usikkert.

Kji-kvadrat-metoden - 2

- Fordeler:
- Bedre filtrering.
- Kan justeres til å ta hensyn til ord-avhengighet.

- Ulemper:
- Meget ad-hoc. Benytter seg av tankegang man vet ikke stemmer helt.
- Trenger en god del global analyse før brukerne drar nytte av den. (Mindre adaptiv?)

LSA I

- Vektor-rom modellen:

	d1	.	.	dn
w1				
w2				
.				
.				
wn				

- Ord-frekvens i dokumenter modelleres som punkter i et n-dimensjonalt rom. Semantisk tilhørighet beregnes fra distanse i vektorrommet.

LSA:

Bruker regresjon (SVD) for å konstruere en avbildning av den opprinnelige matrisen, som skal avdekke en latent semantisk struktur mellom ord og dokumenter.

SVD = PCA.

LSA II

- «Spammete» ord vil havne nær hverandre i vektorrommet fordi de opptrer i like kontekster.
- Opprinnelig en teknikk utviklet for søking i tekst
- Implementert i Apple´s spamfilter.

Noen tanker rundt statistikk-basert software og brukervennlighet

- Statistisk analyse gjøres her uten at noe annet menneske enn brukeren er involvert.
- Det fungerer fordi input og output er forståelig for brukeren.
- Pga mengden og responstid; uaktuelt å involvere statistikere etter implementering.
- Peker fremover; Det kan være mange områder der det er behov for statistisk analyse gjort av ikke-statistikere.
- Som i vanlig software-utvikling er brukervennlighet avgjørende.

Det spillteoretiske aspektet

- Det statistiske filteret er i konkurranse med den menneskelig intelligensen til spammerne.
- I utgangspunktet en lite rettferdig konkurranse, men spillereglene er langt fra like.
- Filterets fordeler; tilpasser seg ordbruken, personlig regelsett
- Filterets ulemper; den underliggende logikken er ikke adaptiv
- Spammerens fordeler; menneskelig intelligens, kan forandre spillestil
- Spammerens ulemper; må kommunisere med mottagere om spesifikke emner, lite økonomisk incentiv

Spammernes mottrekk

- Sende som bilde – Ulempe; eposten vil bli massivt større, noen epost-klienter viser ikke bilder.
- Sende som bilde-link – Fordel; Kan sende med unik id. Ulempe; noen epost-klienter tar ikke html, andre kan filtrere på html-raten.
- Sende kun eksplisitt link. Ulempe; Mindre sannsynlig at mottageren blir interessert.
- Putte inn ekstra tekst (ikke spam-relatert). Fordel; forurenser filteret. Ulempe; forvirrer mottageren.
- Feilskriving av spamrelaterte ord (viagra -> v1ägRá). Ulempe; forvirrer mottageren, vil etterhvert dukke opp i filterets ordliste, filtrering på raten av store bokstaver og spesial-karakterer.

Raffinering av metodene

- Skille mellom ord i tekst og ord i ulike deler av epost-header.
- Se på frekvens av fraser i stedet for ord (CRM114).
- Crawler for å sjekke ut linker i meldingen.
- Ta hensyn til avhengighet mellom ord i kji-kvadrat-metoden eller ved hjelp av Bayesianske nettverk.

Oppsummering/kommentarer

- Ulempe; metodene inneholder 'tweaks' optimert for spammen slik den i programmererens epost-samling under implementering.
- Slike finjusteringer er vanskelig å rettferdiggjøre statistisk.
- Grunnen er at metodene er lagd av programmerere med moderat statistisk bakgrunn.
- Test-resultatene er dermed gjort på ganske ulike bakgrunn, og er vanskelige å sammenligne.
- Den metoden som lettest kan fortolkes fra et statistisk ståsted (naiv Bayes) inneholder antagelser som umulig kan være korrekte.
- Her er det derfor åpning for forbedringer.
- Tankegangen i metodene kan kanskje anvendes andre stedet også?