Informasjonsgjenfinning

INF5820 – H2008

Jan Tore Lønning

Institutt for Informatikk
Universitetet i Oslo

18. september

Outline

Hva er IR?

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
(Manning, Raghavan, Schütze 2008)

Tradisjonelt søk

- En samling dokumenter
- En søkefrase
- Eksakt søk: et dokument er enten relevant eller ikke relevant
- Boolske søk: kombiner søketermer med bruk av AND OR NOT osv.
- Eksempel: en biliotekskatalog,
  Finn alle bøkene med `forfatter: Bjørnson`

Eksakt evaluering

- Anta at:
  - X er mengden av alle dokumenter.
  - A er de dokumentene i X som er relevante for søket
  - B er dokumentene som systemet finner
- Nøkkelbegreper
  - $$\text{Precision: } \frac{|A \cap B|}{B}$$
  - $$\text{Recall: } \frac{|A \cap B|}{A}$$
  - $$\text{F-score (vanlig): } \frac{2PR}{R+P}$$
  - $$\text{F-score (generell): } \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}, \text{ for en} \alpha \in [0,1]$$

Precision and recall

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$P = TP/(TP + FP)$$
$$R = TP/(TP + FN)$$

Invertert indeks

- Alle termene som forekommer i dokumentsamlingen
- Sammen med hver term: id for dokumentene hvor det forekommer
- Evt. også hvor i dokumentet
- Termene lagret alfabetisk for rask gjenfinning
- Hvis dokument-idene har en fast ordning, kan kombinerte søk gå raskt
- Hvor i dokumentet muliggjør søk på flerordsfraser *grønn te*

IR og web

- Med WWW ble IR mye viktigere.
- Vi er helt avhengig av det
- Søk har endret karakter
- Ikke eksakt match som teller—kan være for mange
- Relevans: Få de mest sentrale først!

Rangering

- Precision og recall: Vi sammenlikner alle funnete dokumenter med alle vi skal finne
- Ikke alltid interessert i alle dokumenter, men i ett (eller flere) som gir et svar
- Web-søk
- Anta at vi finner 1000 dokumnter og at 100 av dem er relevante. Forskjell på å få disse først og sist
- Rank-specific precision and recall, jfr. J& M figure 23.4
- Precision-recall graf

Glatting og interpolering

- Ser vi på bare et søk, kan vi få en graf som:
  - Er hakkete
  - Kan bli stigende
- Interpolering: Ta maksimum av fremtidige punkter.
- Average: Ta gjennomsnitt av flere punkter (= ett tall)
- "Interpolated average": En kombinasjon,
  - Gjennomsnittet for alle punkter med inntil 10% recall, alle med inntil 20% recall osv.
  - Deretter interpolér
- For å sammenlikne ulike IR-algoritmer må vi se på flere søk og ta gjennomsnitt.

Forutsetninger:

- Et dokument blir sett på som en “bag” av termer.
- (Ingen forskjell på `The box is in the pen` og `The pen is in the box`)
- Et søkeuttrykk blir også sett på som en “bag” av termer
- Likhet mellom dokumenter måles om de har samme termprofil

Example

Dokumentnr.	antall *bass*	antall *fishing*
Dok1	10	2
Dok2	10	20
Dok3	2	10
Dok4	25	10

- Hvilken av de andre likner mest på Dok 1?

- Liknende dokumenter har liknende relativ fordeling mellom termene

Litt vektorregning

Repetisjon:

- Elementær geometri: sinus, cosinus, tangens
- Vektorer og skalarprodukt
- Lengden av en vektor
- Skalarprodukt og cosinus

Vektorlikhet

Hva skal vi se på?

- Differanse, eller
- Skalarprodukt?

Differanse

- Lengden av avstandsvektoren
- Manhatten: $\sum_{i=1}^{N} |x_i - y_i|$
- Euklidsk: $\sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$

Produkt

- $$\text{sim}_{\text{dot-product}}(\vec{v}, \vec{w}) = \vec{v} \cdot \vec{w} = \sum_{i=1}^{N} v_i \times w_i$$
- $$\text{sim}_{\text{cosine}}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}|\,|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}}$$
- =1 hvis de peker i samme retning

Produkt og differense

For normaliserte vektorer gir produkt og differanse samme rangering.

IDF

- Skal alle termer telle like mye?
- Idé: Termer som forekommer i få dokumenter er karakteristiske for de dokumentene hvor de forekommer
- N er antall dokumenter og n_i er de hvor term i forekommer
- IDF, **inverse document frequency**: $\frac{N}{n_i}$
- $$\text{idf}_i = \log\left(\frac{N}{n_i}\right)$$
- **td-idf**-vekting: $w_{i,j} = \text{tf}_{i,j} \times \text{idf}_i$
- der $\text{tf}_{i,j}$ antall forekomster av term i i dokumemt j
- Sentralt i flere anvendelser, som summering.

Relasjonen mellom IR og NLP

- Ser en stort på det er IR en del av NLP, men
- De har utviklet seg som to uavhengige disipliner, ulikt fokus
- IR: Vektorer, statistikk
- NLP/datalingvistikk: grammatikk, parsing
- To typer tilnærminger:
  - Bruk av lingvistisk kunnskap i IR
  - Økende bruk av statistikk og vektorer i NLP

Hva er en term?

- Et ord — fullform, leksem, noe annet?
- Tradisjonelt: "stemming": *kaste, kaster, kastet, kast* blir alle kuttet til *kast*
- Hva med *spiser, spiste, spist* skal de gå til *spis*?
- Eller hva med *glemmer, glemte*? *går, gikk*?
- En stemmer er "the poor man's lemmatizer":
  - Trenger ikke ordbok
  - Ikke perfekt
- Stemming kan bedre "recall", men
- Gi dårligere presisjon. Slår f.eks. sammen substantiv og verb for *løpe*

Andre bidrag fra (data)lingvistikk til IR?

- Eksperimentering, men ikke entydige resultat:
- Lemmatisering
- Tagging
- WSD
- (Grammatiske) relasjoner, jfr. forskjell på `The box is in the pen` og `The pen is in the box` (Kommer!)

Oppsummering

- Precision og recall
- Invertert indeks
- Forutsetningene som ligger under vektormodellen
- Likhet mellom vektorer: cosinus og skalarprodukt
- idf
- Stemming

Noen kilder

- Ny bok om IR: Manning, Raghaven, Schütze, *Introduction to Information Retrieval*, 2008
- Stor utbredelse de senere år: Baeza-Yates, Ribeiro-Neto, *Modern Information Retrieval*, 1999
- Flere temaer i INF5820: Manning, Schütze, *Foundations of Statistical Language Processing*, 1999