Capitole Speciale de Informatică

Curs 5: Extragerea informațiilor prin feedback de relevanță. Metode probabiliste de extragere a informațiilor

25 octombrie 2018

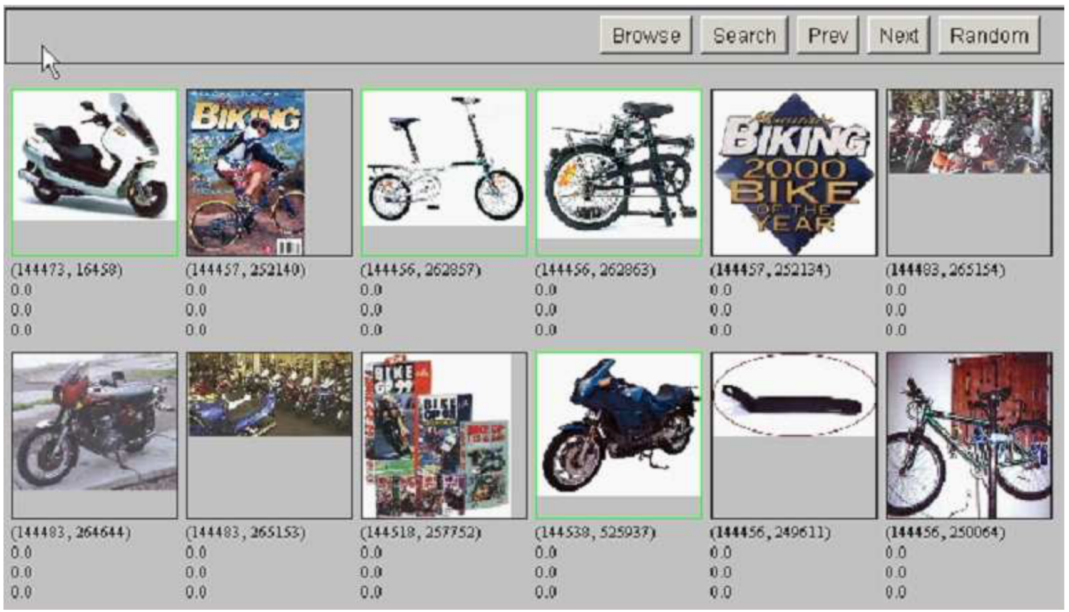
1. Utilizatorul comunică o cerere simplă de informare.
2. Sistemul de IR returnează o mulțime inițială de rezultate de căutare.
3. Utilizatorul marchează unele din rezutatele căutării ca *relevante*, și altele ca *irelevante*.
4. Sistemul calculează o reprezentare mai bună a cererii de căutare, pe baza feedback-ului primit de la utilizator.
5. Sistemul afișează o colecție actualizată de rezultate de căutare.

Pașii 3-5 pot fi repetați de mai multe ori.

Motivație: Uneori, este dificil să se formuleze o cerere de informare bună pentru o nevoie de informare. Prin feedback interactiv, sistemul poate rafina cererea de căutare a utilizatorului

Sistemul demonstrativ (momentan offline)
`http://nayana.ece.ucsb.edu/imsearch/imsearch.html`

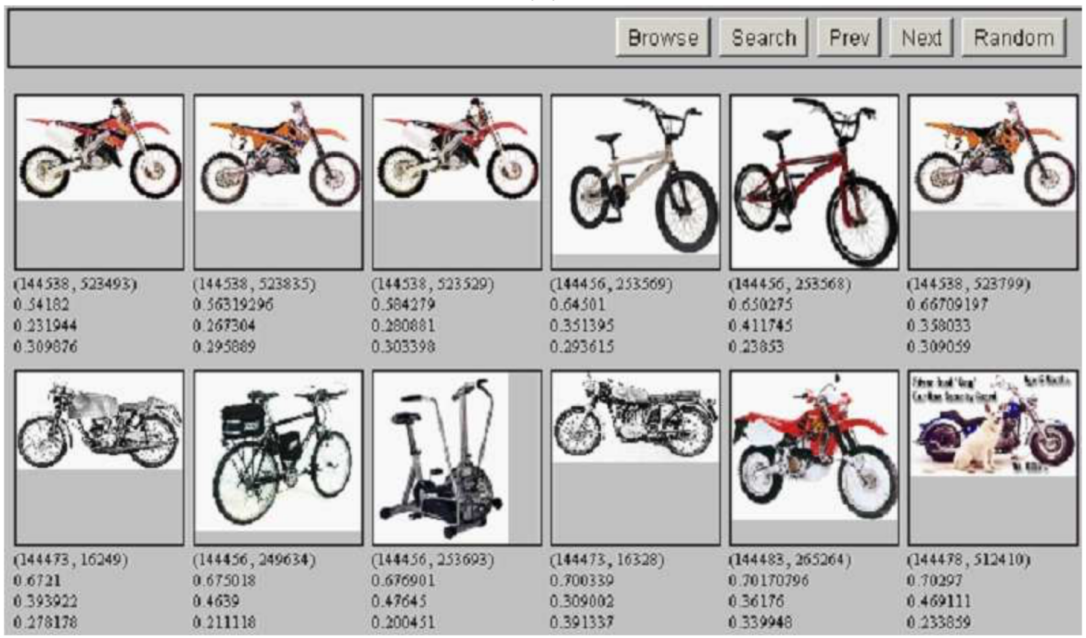
- interacțiune declanșată de cererea inițială `bike`



După selectarea a 4 imagini ca relevante, sistemul afișează o colecție actualizată de rezultate (vezi slide-ul următor):

Extragerea informaţiilor prin feedback de relevanţă

Exemplu ilustrat de căutare în o colec ti de imagini (2)



OBSERVAŢIE: După feedback, rezultatele sunt mult mai bune.

(a) Query: New space satellite applications

(b)
\+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
\+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
\+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

(c)

2.074 new 15.106 space
30.816 satellite 5.660 application
5.991 nasa 5.196 eos
4.196 launch 3.972 aster
3.516 instrument 3.446 arianespace
3.004 bundespost 2.806 ss
2.790 rocket 2.053 scientist
2.003 broadcast 1.172 earth
0.836 oil 0.646 measure

(d)

* 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan

* 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer

3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own

4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit

* 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies

6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use

7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers

8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Idee de bază: După ce utilizatorul primește colecția de răspunsuri C la cererea q, și marchează submulțimea de documente $C_r \subseteq C$ ca relevante, și $C_{nr} \subseteq C$ ca nerelevante, vrem să rafinăm cererea q în o cerere q_{opt} astfel încât diferența

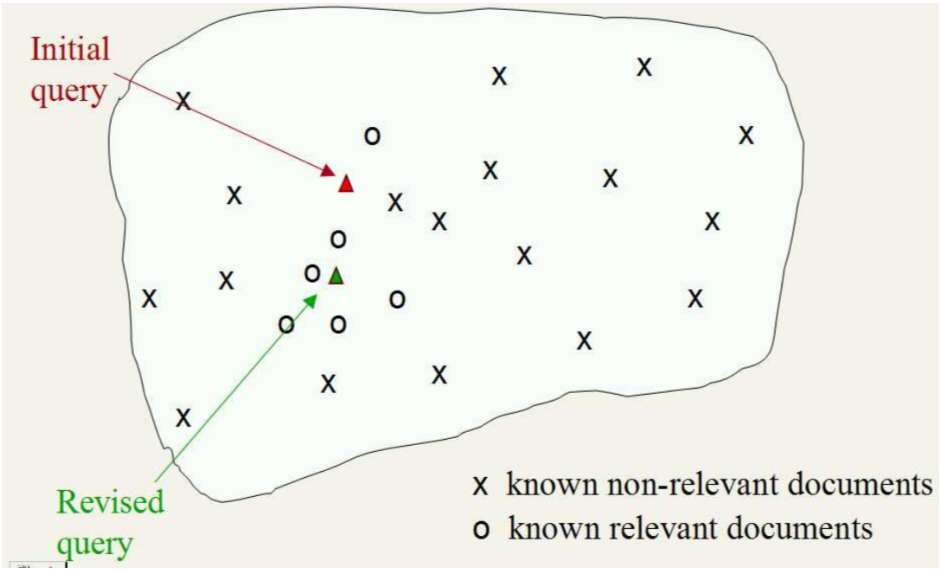
$$\text{sim}(\vec{q}, C_r) - \text{sim}(\vec{q}, C_{nr})$$

ia valoarea maximă pentru $\vec{q} = \vec{q}_{opt}$.

- $\text{sim}(\vec{q}, C_r)$ măsoară gradul de similaritate dintre cererea q și colecția C_r de documente relevante
- $\text{sim}(\vec{q}, C_{nr})$ măsoară gradul de similaritate dintre cererea q și colecția C_{nr} de documente nerelevante

Dacă se folosește similaritatea cosinusoidală, rezultă că

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{d \in C_r} \vec{d} - \frac{1}{|C_{nr}|} \sum_{d \in C_{nr}} \vec{d}$$



Dupa ce utilizatorul a marcat unele răspunsuri ca relevante și altele ca nerelevante, vectorul inițial de interogare și-a schimbat poziția.

- Algoritmul Rocchio a fost propus in 1971 și folosit prima oară de sistemul SMART de extragere a informațiilor.
- De obicei, se folosește următoarea variație a formulei de calcul pentru modificarea de către feedback a cererii ($\vec{q}_m$):

$$\vec{q}_m = \alpha \cdot \vec{q}_0 + \beta \cdot \frac{1}{|D_r|} \sum_{d \in D_r} \vec{d} - \gamma \cdot \frac{1}{|D_{nr}|} \sum_{d \in D_{nr}} \vec{d}$$

unde $\alpha, \beta, \gamma \in \mathbb{R}$ sunt greutăți pentru gradele de importanță ale cererii inițiale (α), colecției de documente relevante (β) și colecției de documente nerelevante (γ)

- de obicei, $0 \leq \gamma < \beta$
- valori rezonabile sunt $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.15$

Feedback de relevanță probabilist

Idee: În loc să modificăm cererea, construim un clasificator care deosebește mai bine răspunsurile relevante de cele nerelevante.

Modelul probabilist Bayes naiv calculează pentru fiecare termen t următoarele estimări:

- $\hat{P}(x_t = 1 \mid R = 1) = \frac{|VR_t|}{|VR|}$: probabilitatea că t apare în un document relevant
- $\hat{P}(x_t = 1 \mid R = 0) = \frac{\mathrm{df}_t - |VR_t|}{N - |VR|}$: probabilitatea că t apare în un document irelevant

unde: df_t este frecvența de document a lui t (nr. de documente din colecție în care apare t); VR este mulțimea reală de documente relevante; VR_t este submulțimea din VR în care apare t; N este numărul total de documente.

- Estimările de probabilitate $\hat{P}(x_t = 1 \mid R = 1)$ și $\hat{P}(x_t = 1 \mid R = 0)$ se folosesc pentru a recalcula greutățile termenilor din documentele relevante.
- Algoritmul de feedback probabilist folosește doar valori statistice despre destribuția termenilor in documente considerate relevant. Nu se reține informații despre cererea originală q.

Feedback de relevanță pentru căutarea web

- Inițial, motorul de căutare Excite a oferit suport pentru feedback de relevanță
  - ulterior, s-a renunțat la această capabilitate, din lipsă de interes din partea utilizatorilor
- Unele motoare de permit căutarea de pagini similare cu anumită pagină web selectată de utilizator.
- O metodă de feedback indirect de relevanță este monitorizarea hyperlink-urilor urmate de utilizator pentru a accesa informații. Această tehnică exploatează structura link-urilor din rețeaua web.

Dată fiind o cerere de informare q și un document d, se estimează

- $P(R = 1 \mid d, q)$: probabilitatea ca documentul p să fie relevant pentru cererea q.
- $P(R = 1 \mid d, q)$: probabilitatea ca documentul p să fie nerelevant pentru cererea q.

și se ordonează documentele descrescător după valoarea estimată a lui $P(R = 1 \mid d, q)$.

Dată fiind o cerere de informare q și un document d, se estimează

- $P(R = 1 \mid d, q)$: probabilitatea ca documentul p să fie relevant pentru cererea q.
- $P(R = 1 \mid d, q)$: probabilitatea ca documentul p să fie nerelevant pentru cererea q.

și se ordonează documentele descrescător după valoarea estimată a lui $P(R = 1 \mid d, q)$.

Regula de decizie Bayes optimală impune să se returneze doar documente pt. care probabilitatea de a fi relevante este mai mare decât cea de a fi nerelevante:

d este relevant dacă și numai dacă $P(R = 1 \mid d, q) > P(R = 0 \mid d, q)$

Fie C_1 costul nereturnării unui document relevant, și C_1 costul returnării unui document nerelevant.

Principiul probabilist de gradare cu costuri de extragere prevede că, dacă d este un document a.î.

$$C_0 \cdot P(R = 0 \mid d, q) - C_1 \cdot P(R = 1 \mid d, q) \leq$$
$$C_0 \cdot P(R = 0 \mid d', q) - C_1 \cdot P(R = 1 \mid d', q)$$

pentru orice document d' care are urma să fie extras, atunci d este documentul următor care se extrage.

Modelul folosit în mod tradițional pt. implementarea principiului probabilist de gradare.

- **Binar** provine de la boolean, adică reprezentarea vectorială a unui document d este $\vec{d} = (x_1, \ldots, x_M)$ unde $x_i = 1$ dacă $t_i \in d$ și $x_i = 0$ în caz contrar.
- **Independență**: aparițiile termenilor în un document sunt modelate independent; nu sunt modelate relații între termeni.
- BIM presupune că relevanța fiecărui document este independentă de relevanța altui document.
- Convenții de notație: $P(\vec{d} \mid R = 1, \vec{q})$ (resp. $P(\vec{d} \mid R = 0, \vec{q})$): probab. ca $\vec{d}$ să fie returnat (extras), știind că $\vec{d}$ este relevant (resp. irelevant) pentru cererea q; $P(R = 1 \mid \vec{q})$ (resp. $P(R = 0 \mid \vec{q})$): probab. ca un document din colecție să fie relevant (resp. irelevant) pt. $\vec{q}$.

$$\text{Știm că } P(R=1 \mid \vec{d},\vec{q}) = \frac{P(\vec{d} \mid R=1,\vec{q}) \cdot P(R=1 \mid \vec{q})}{P(\vec{d} \mid \vec{q})}$$

$$P(R=0 \mid \vec{d},\vec{q}) = \frac{P(\vec{d} \mid R=0,\vec{q}) \cdot P(R=0 \mid \vec{q})}{P(\vec{d} \mid \vec{q})}$$

$$P(R=1 \mid \vec{d},\vec{q}) + P(R=0 \mid \vec{d},\vec{q}) = 1$$

și vrem să ordonăm documentele în ordine descrescătoare a lui $P(R=1 \mid \vec{d},\vec{q})$.

Observații

Dacă $0 \leq x < y < 1$ atunci $\frac{x}{1-x} < \frac{y}{1-y} \Rightarrow$ este suficient să ordonăm documentele descrescător după

$$\frac{P(R=1 \mid \vec{d},\vec{q})}{1-P(R=1 \mid \vec{d},\vec{q})} = \frac{P(R=1 \mid \vec{d},\vec{q})}{P(R=0 \mid \vec{d},\vec{q})} = \frac{P(R=1 \mid \vec{q})}{P(R=0 \mid \vec{q})} \cdot \frac{P(\vec{d} \mid R=1,\vec{q})}{P(\vec{d} \mid R=0,\vec{q})}$$

Modelul probabilist de independență binară (BIM)

Ordonează documentele d descrescător după

$$\underbrace{\frac{P(R=1 \mid \vec{q})}{P(R=0 \mid \vec{q})}}_{\text{parte constantă}} \cdot \frac{P(\vec{d} \mid R=1, \vec{q})}{P(\vec{d} \mid R=0, \vec{q})}$$

adică după

$$\frac{P(\vec{d} \mid R=1, \vec{q})}{P(\vec{d} \mid R=0, \vec{q})} = \prod_{i=1}^{M} \frac{P(x_i \mid R=1, \vec{q})}{P(x_i \mid R=1, \vec{q})}$$

$$= \prod_{i:x_i=1} \frac{P(x_i=1 \mid R=1, \vec{q})}{P(x_i=1 \mid R=1, \vec{q})} \cdot \prod_{i:x_i=0} \frac{P(x_i=0 \mid R=1, \vec{q})}{P(x_i=0 \mid R=1, \vec{q})}$$

Modelul probabilist de independență binară (BIM)

Fie $p_{t_i} := P(x_i = 1 \mid R = 1, \vec{q})$ și $u_{t_i} = P(x_i = 1 \mid R = 0, \vec{q})$. Tabelul de mai jos ilustrează semnificația acestor probabilități:

	document	relevant ($R = 1$)	nerelevant ($R = 0$)
termen prezent	$x_t = 1$	p_t	u_t
termen absent	$x_t = 0$	$1 - p_t$	$1 - u_t$

Rezultă că documentele se ordonează descrescător după

$$\frac{P(\vec{d} \mid R = 1, \vec{q})}{P(\vec{d} \mid R = 0, \vec{q})} \cdot \prod_{i:x_i=q_i=1} \frac{p_{t_i}}{u_{t_i}} \cdot \prod_{i:x_i=0,q_i=1} \frac{1 - p_{t_i}}{1 - u_{t_i}}$$

$$= \frac{P(\vec{d} \mid R = 1, \vec{q})}{P(\vec{d} \mid R = 0, \vec{q})} \cdot \prod_{i:x_i=q_i=1} \frac{p_{t_i}(1 - u_{t_i})}{u_{t_i}(1 - p_{t_i})} \cdot \prod_{i:q_i=1} \frac{1 - p_{t_i}}{1 - u_{t_i}}$$

Factorii roșii nu depind de document $\Rightarrow$ trebuie să ordonăm descrescător după factorul albastru, sau după logaritmul lui (Retrieval Status Value):

- $RSV_d := \log \prod_{i:x_i=q_i=1} \frac{p_{t_i}(1 - u_{t_i})}{u_{t_i}(1 - p_{t_i})} = \sum_{i:x_i=q_i=1} \log \frac{p_{t_i}(1 - u_{t_i})}{u_{t_i}(1 - p_{t_i})}$

$RSV_d = \sum_{i:x_i=q_i=1} c_i$ unde

$$c_i = \log \frac{p_{t_i}(1-u_{t_i})}{u_{t_i}(1-p_{t_i})} = \log \frac{p_{t_i}}{1-p_{t_i}} + \log \frac{1-u_{t_i})}{u_{t_i}(1-p_{t_i})}$$

- $c_t = 0$ dacă t are aceeași șansă să apară în un document relevant ca și în un document nerelevant
- $c_t > 0$ dacă t are șansă mai mare să apară în un document relevant.

Întrebare: cum se pot estima valorile lui c_t?

Modelul probabilist de independență binară (BIM)

Estimarea teoretică a valorilor lui c_t pentru o cerere q și colecție D

Dacă D are N documente, S este numărul total de documente relavante pt q, s este nr. numărul total de documente relavante pt q care-l conțin pe t, atunci avem situația din tabelul de mai jos:

	documente	relevant	nerelevant	Total
Termen prezent	$x_t = 1$	s	$\mathrm{df}_t - s$	df_t
Termen absent	$x_t = 0$	$S - s$	$N - \mathrm{df}_t) - (S - s)$	$N - \mathrm{df}_t$
	Total	S	$N - S$	N

Rezultă că $p_t = s/S$, $u_t = (\mathrm{df}_t - s)/(N - S)$ și

$$c_t = \log \frac{s/(S-s)}{(\mathrm{df}_t - s)/((N - \mathrm{df}_t) - (S - s))}$$

Pentru a evita valori nule, se calculează

$$\hat{c}_t = K(N, \mathrm{df}_t, S, s) = \log \frac{(s + \frac{1}{2})/(S - s + \frac{1}{2})}{(\mathrm{df}_t - s + \frac{1}{2})/((N - \mathrm{df}_t) - S + s + \frac{1}{2})}$$

Presupunând ca procentul documentelor relevante din colecție este f. mic, putem aproxima $u_t \approx \mathrm{df}_t/N$ și

$$\log \frac{1-u_t}{u_t} = \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t} \approx \log \frac{N}{\mathrm{df}_t}$$

Există mai multe metode de aproximare a cantității p_t:

1. Unii cercetători propun utilizarea frecvenței de termeni în documentele relevante cunoscute (dacă se cunosc unele).
2. Alți cercetători propun utilizarea constantei $p_t = 0.5$
3. Greiff (1998) propune $p_t = \frac{1}{3} + \frac{2}{3}\mathrm{df}_t/N$

Bibliografie

1. *Relevance feedback and query expansion* (Cap. 9) și *Probabilistic information retrieval* (Cap. 11) din Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: *An Introduction to Information Retrieval*. Ediție online (c) 2009 Cambridge UP.
   `http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf`