Pristupi automatskoj predmetnoj klasifikaciji tekstualnih Web-stranica

Koraljka Golub,
Knowledge Discovery and Digital Libraries Research Group (KnowLib)
Odsjek za informacijsku tehnologiju, Sveučilište u Lundu, Švedska

SZI seminar, Zagreb, 5. i 6.11.2004.

Sadržaj

1. Uvod
- Što je automatska klasifikacija
- Potreba za automatskom predmetnom klasifikacijom (AK)
- Primjeri primjene AK

2. Postojeći pristupi AK
- Kategorizacija teksta
- Grupiranje dokumenata
- Klasifikacija dokumenata
- Kombinirani pristup

3. Glavna pitanja

Što je *automatska klasifikacija*

- klasifikacija
  - širi pojam:
    - razvrstavanje u klase, kategorije, skupine
  - uži pojam:
    - grupiranje dokumenata prema predmetu
- neki srodni pojmovi:
  - kategorizacija (categorization)
  - grupiranje (clustering)
  - predmetno označivanje (subject indexing)
- **automatska predmetna klasifikacija (AK)**
  - strojno (nemanuelno) grupiranje dokumenata po predmetu

Potreba za AK i primjena...

- potreba za automatskom klasifikacijom Web-dokumenata
  - sve veći broj dokumenata na WWW-u
  - visoki troškovi manuelne predmetne obrade
- razne mogućnosti primjene
  - grupiranje novinskih članaka po temi, ili automatsko personalizirano slanje vijesti
  - filtriranje neželjenog sadržaja (npr. govor mržnje) u preglednicima za Internet (Internet browsers), ili *spam* poruka u programima za e-poštu

...Primjeri primjene AK

- grupiranje rezultata pretraživanja po temama (npr. u Web-tražilici), npr.
  - http://www.clusty.com, http://www.objectssearch.com/
  - http://www.kartoo.com/
- grupiranje Web-stranica u tematske skupine za prebiranje (browsing), tj. automatska izrada tematskih portala (subject gateways), općih ili specijaliziranih, npr.
  - http://engine-e.lub.lu.se/
  - http://www.it.lth.se/knowlib/demos.htm

AK tekstualnih Web-stranica

- naša tema: automatska klasifikacija tekstualnih Web-stranica, u različite vrste predmetnih kategorija
- tekstualne Web-stranice
  - hiperveze (bibliometrijske metode)
  - *anchors*
  - metapodaci
  - strukturalne oznake (<h1> važan tekst</h1>)

ALI:
- vrlo heterogene
  - kratke
  - metapodaci nekonzistentni, netočni
  - strukturalne oznake nekonzistentne
  - naslovi opći ("Home page")

Postojeći rezultati

- raznoliki
  - 50-60% ispravno klasificiranih dokumenata
  - neki govore o 90% ispravno klasificiranih dokumenata
- teško usporedivi
  - različite mjere evaluacije
    - problem evaluacije: relativnost predmeta, djelomična točnost
  - različite zbirke dokumenata za testiranje (test collections)
  - različita primjena

Sadržaj

1. ✓ Uvod
   - ✓ Što je automatska klasifikacija
   - ✓ Potreba za automatskom predmetnom klasifikacijom (AK)
   - ✓ Primjeri primjene AK

2. Postojeći pristupi AK
   - Kategorizacija teksta
   - Grupiranje dokumenata
   - Klasifikacija dokumenata
   - Kombinirani pristup

3. Glavna pitanja

Tri osnovna pristupa

- *Text categorization*
- *Document clustering*
- *Document classification*
- U čemu se pristupi razlikuju?
  - metodologija
  - znanstvena tradicija
  - svojstva kategorija
  - zbirke dokumenata za testiranje
  - metode evaluacije
  - primjena

Kategorizacija teksta (KT) (text categorization)...

- tri osnovna koraka:

1) manuelna kategorizacija određenog broja dokumenata za učenje (training documents) u postojeće, manuelno izrađene, kategorije

2) "učenje" algoritama o svojstavima kategorija, na temelju tih dokumenata za učenje:
   - 2.1) vektorski prikaz svake kategorije (vektor1) i vektorski prikaz svakog dokumenta (vektor2) na temelju, npr. frekvencije pojedinih riječi u kategoriji i u dokumentu
     - učenje o svojstvima pojedinih kategorija temelji se na algoritmima za strojno učenje
   - 2.2) utvrđivanje stupnja razlike između tih vektora (vektor1 i vektor2), npr. mjerom za kosinus kuta (cosine measure) između njih, da bi se utvrdila njihova sličnost
     - što imaju više zajedničkih riječi, to je razlika manja

3) automatska klasifikacija novih dokumenata

...Kategorizacija teksta

- strojno učenje
- korišteni sustavi predmetnih kategorija obično nemaju mehanizme za "kontrolu" nad polisemijom i ostalim svojstvima prirodnog jezika problematičnima pri pretraživanju
- zbirke: Reuters, OHSUMED, baza podataka US Patent, 20 Newsgroups DataSet
  - za Web-dokumente: WebKB, postojeći direktoriji (Yahoo!)
- preciznost i odziv i ostale mjere iz informacijskog pretraživanja
- primjena u operativnim informacijskim sustavima
  - komercijalni sustavi, npr.

  Thunderstone Web Site Catalog http://search.thunderstone.com/texis/websearch/

Grupiranje dokumenata (GD) (document clustering)...

- za razliku od kategorizacije teksta,
  - ne obuhvaća prethodno određene kategorije
  - ne obuhvaća dokumente za učenje
- grupe sličnih dokumenata, kao i njihovi nazivi, dobivaju se na temelju dokumenata koje se nastoji grupirati
  - osnovni problem: nazivanje grupa i definiranje odnosa među njima
- dokumenti su predstavljeni kao vektori, a zatim se uspoređuju raznim mjerama sličnosti (similarity measures)
- slični dokumenti svrstavaju se algoritmima u automatski generirane predmetne skupine

...Grupiranje dokumenata

- informacijsko pretraživanje (information retrieval)
- predmetne kategorije automatski su generirane
  - velik problem su automatski generirani nazivi kategorija
  - vrlo je teško automatski generirati odnose među pojmovima (hijerarhijske, sinonimne itd.)
  - ne postoji kontrola na problemima prirodnog jezika
- zbirke: TREC, INEX (u razvoju)
- preciznost i odziv i ostale mjere iz informacijskog pretraživanja
- primjena:
  - grupiranje rezultata pretraživanja: http://www.clusty.com
  - prebiranje: http://engine-e.lub.lu.se

Klasifikacija dokumenata (KD) (document classification)...

- proces:
  - odabir termina iz dokumenta
  - rangiranje termina po vjerojatnoj relevantnosti
  - usporedba odabranih termina s terminima u kontroliranom rječniku
  - uglavnom se ne koristi vektorski pristup
- najčešće ne koristi složene algoritme poput onih primjenjivanih u kategorizaciji i grupiranju
- fokus istraživanja su više operativni informacijski sustavi, a manje eksperimenti u kontroliranim uvjetima

...Klasifikacija dokumenata...

- knjižnična znanost
- manuelno oblikovani sustavi za označivanje poput klasifikacijskih sustava, u čije se klase ili kategorije dokumenti svrstavaju
  - ovi sustavi osiguravaju nadzor nad problemima sinonimije i polisemije prirodnog jezika
- zbirke
  - *harvested* Web-dokumenti, bibliografski zapisi Web-dokumenata
- evaluacija
  - korisnici, mjere iz informacijskog pretraživanja

...Klasifikacija dokumenata...

- primjeri
  - GERHARD http://www.gerhard.de/
    - UDK
  - Scorpion http://orc.rsch.oclc.org:6109/
    - DDK, svaki dokument poslan kao upit u DDK bazu
    - za grupiranje rezultata koristili grupiranje dokumenata (clustering)
    - OCLC - razvoj alata za AK: WordSmith, FAST (Faceted Application of Subject Terminology)
  - "All" Engineering http://eels.lub.lu.se/ae/index.html i Engine-e http://engine-e.lub.lu.se/
    - tezaurus Ei (Engineering Index) i DDK

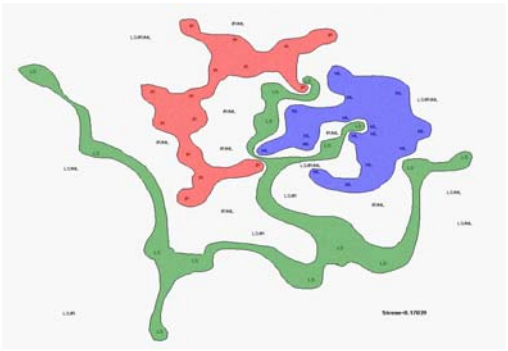
Usporedba (najčešći slučajevi)

korištene metode	*KT*	*GD*	*KD*
indeksiranje – uklanjanje stop riječi – korjenovanje – odabir najboljih riječi i/ili izraza •*term weighting*	✓	✓	✓
algoritmi	ML	IR	KD
pretvaranje dokumenata u vektore	✓	✓	--
eksploatiranje raznih svojstava Web-dokumenata	✓	✓	✓

Kombinirani pristup

- korisnost kontroliranih rječnika sve se više uvažava
  - → u okviru kategorizacije teksta ili grupiranja dokumenata koriste se (manuelno oblikovani) kontrolirani rječnici
  - primjeri korištenja kategorija iz, npr., NorthernLight, Yahoo!, i drugih tražilica
  - primjeri korištenja LKK i MeSH za grupiranje rezultata pretraživanja
  - primjeri koji potvrđuju važnost dobrih hijerarhijskih struktura za poboljšanje točnosti automatske klasifikacije
- ovaj pristup upućuje na korisnost razmjene ideja između triju glavnih pristupa
  - hipoteza: razmjena ideja, metoda, pristupa itd. je korisna

Razmjena ideja danas



Bibliographic coupling map based on 62 selected papers per each community with coloured groupings.

Sadržaj

1. ✓ Uvod
   - ✓Što je automatska klasifikacija
   - ✓ Potreba za automatskom predmetnom klasifikacijom (AK)
   - ✓ Primjeri primjene AK
2. ✓ Postojeći pristupi AK
   - ✓ Kategorizacija teksta
   - ✓ Grupiranje dokumenata
   - ✓ Klasifikacija dokumenata
   - ✓ Kombinirani pristup
3. Glavna pitanja

Glavna pitanja...

- problem odabira termina iz dokumenta
  - brojni pristupi
  - u TK i GD problem vektorskog prostora u kojem se riječi i dokumenti predstavljaju kao međusobno neovisni, bez uvažavanja konteksta (broji se samo frekvencija)
    - teorijski neopravdana manipulacija vektorima, npr. korištenje kosinus mjere za dobivanje sličnosti dokumenata
- problem svojstava kategorija
  - u GD se dobivaju automatski
    - "Automatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen, Dumais 2000)
    - u GD dodatan problem: nestabilnost strukture za prebiranje
  - u TK i GD nema kontrole nad prirodnim jezikom

...Glavna pitanja...

- predmetno prebiranje se puno koristi, ali istraživanja su danas usmjerena na tražilice (search engines)
- uloga kontroliranih rječnika sve prepoznatljivija
  - prebiranje
  - prilagođavaju se elektroničkom okruženju (Electronic Dewey, FAO Agricultural Thes...)
  - važnost hijerarhijske struktura za točniji proces automatizacije ("All" Engineering, nekoliko istraživanja iz ostalih pristupa)

...Glavna pitanja

- problem evaluacije
  - mjere informacijskog pretraživanja
  - kvaliteta predmetnog označivanja/klasifikacije
    - koja je predmetna oznaka točna?
    - dubina (specificity), potpunost (exhaustivity)
    - primjerenost korisniku

Hvala na pažnji!

Pitanja i komentari?

koraljka.golub@it.lth.se
http://www.it.lth.se/knowlib/