Sažeci, klasifikacija i klasterovanje

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Kako predstaviti rezultate korisniku?

- Najčešće: kao listu – „10 plavih linkova“

Kako predstaviti rezultate korisniku?

- Najčešće: kao listu – „10 plavih linkova“
- Kako opisati svaki dokument u listi?

Kako predstaviti rezultate korisniku?

- Najčešće: kao listu – „10 plavih linkova“
- Kako opisati svaki dokument u listi?
- Ovaj opis je presudan

Kako predstaviti rezultate korisniku?

- Najčešće: kao listu – „10 plavih linkova“
- Kako opisati svaki dokument u listi?
- Ovaj opis je presudan
- Korisnik može da odredi relevantne pogotke na osnovu opisa

Kako predstaviti rezultate korisniku?

- Najčešće: kao listu – „10 plavih linkova“
- Kako opisati svaki dokument u listi?
- Ovaj opis je presudan
- Korisnik može da odredi relevantne pogotke na osnovu opisa
- Ne mora da klikne na sve dokumente sekvencijalno

Opis dokumenta u rezultatu

- Najčešće: naslov, URL, neki metapodaci ...

Opis dokumenta u rezultatu

- Najčešće: naslov, URL, neki metapodaci . . .
- . . . i sažetak

Opis dokumenta u rezultatu

- Najčešće: naslov, URL, neki metapodaci . . .
- . . . i sažetak
- Kako da „izračunamo“ sažetak?

Sažeci

- Dve osnovne vrste: (1) statički (2) dinamički

Sažeci

- Dve osnovne vrste: (1) statički (2) dinamički
- Statički sažetak dokumenta je uvek isti bez obzira na upit kojim je dokument pronađen

Sažeci

- Dve osnovne vrste: (1) statički (2) dinamički
- Statički sažetak dokumenta je uvek isti bez obzira na upit kojim je dokument pronađen
- Dinamički sažeci su zavisni od upita. Pokušavaju da objasne zašto je dokument pronađen za baš taj upit

Statički sažeci

- Obično je statički sažetak podskup dokumenta

Statički sažeci

- Obično je statički sažetak podskup dokumenta
- Jednostavna heuristika: prvih 50-tak reči u dokumentu

Statički sažeci

- Obično je statički sažetak podskup dokumenta
- Jednostavna heuristika: prvih 50-tak reči u dokumentu
- Nešto složenija: izvući iz dokumenta skup „ključnih“ rečenica

Statički sažeci

- Obično je statički sažetak podskup dokumenta
- Jednostavna heuristika: prvih 50-tak reči u dokumentu
- Nešto složenija: izvući iz dokumenta skup „ključnih“ rečenica
  - jednostavne NLP (natural language processing) heuristike za ocenjivanje svake rečenice

Statički sažeci

- Obično je statički sažetak podskup dokumenta
- Jednostavna heuristika: prvih 50-tak reči u dokumentu
- Nešto složenija: izvući iz dokumenta skup „ključnih“ rečenica
  - jednostavne NLP (natural language processing) heuristike za ocenjivanje svake rečenice
  - sažetak je sastavljen od najbolje rangiranih rečenica

Statički sažeci

- Obično je statički sažetak podskup dokumenta
- Jednostavna heuristika: prvih 50-tak reči u dokumentu
- Nešto složenija: izvući iz dokumenta skup „ključnih“ rečenica
  - jednostavne NLP (natural language processing) heuristike za ocenjivanje svake rečenice
  - sažetak je sastavljen od najbolje rangiranih rečenica
  - pristup zasnovan na mašinskom učenju

Statički sažeci

- Obično je statički sažetak podskup dokumenta
- Jednostavna heuristika: prvih 50-tak reči u dokumentu
- Nešto složenija: izvući iz dokumenta skup „ključnih“ rečenica
  - jednostavne NLP (natural language processing) heuristike za ocenjivanje svake rečenice
  - sažetak je sastavljen od najbolje rangiranih rečenica
  - pristup zasnovan na mašinskom učenju
- Najsloženije: složeni NLP sa sintezu/generisanje sažetka

Statički sažeci

- Obično je statički sažetak podskup dokumenta
- Jednostavna heuristika: prvih 50-tak reči u dokumentu
- Nešto složenija: izvući iz dokumenta skup „ključnih“ rečenica
  - jednostavne NLP (natural language processing) heuristike za ocenjivanje svake rečenice
  - sažetak je sastavljen od najbolje rangiranih rečenica
  - pristup zasnovan na mašinskom učenju
- Najsloženije: složeni NLP sa sintezu/generisanje sažetka
  - za većinu aplikacija nije još dovoljno upotrebljivo

Dinamički sažeci

- Prikazati jedan ili više „prozora“ ili fragmenata iz dokumenta koji sadrže termove iz upita

Dinamički sažeci

- Prikazati jedan ili više „prozora“ ili fragmenata iz dokumenta koji sadrže termove iz upita
- Generišu se u skladu sa rangiranjem

Dinamički sažeci

- Prikazati jedan ili više „prozora“ ili fragmenata iz dokumenta koji sadrže termove iz upita
- Generišu se u skladu sa rangiranjem
- Posebno dobri fragmenti: gde se traženi termovi pojavljuju kao fraza

Dinamički sažeci

- Prikazati jedan ili više „prozora“ ili fragmenata iz dokumenta koji sadrže termove iz upita
- Generišu se u skladu sa rangiranjem
- Posebno dobri fragmenti: gde se traženi termovi pojavljuju kao fraza
- Posebno dobri fragmenti: gde se traženi termovi pojavljuju zajedno u malom prozoru (na malom prostoru)

Dinamički sažeci

- Prikazati jedan ili više „prozora“ ili fragmenata iz dokumenta koji sadrže termove iz upita
- Generišu se u skladu sa rangiranjem
- Posebno dobri fragmenti: gde se traženi termovi pojavljuju kao fraza
- Posebno dobri fragmenti: gde se traženi termovi pojavljuju zajedno u malom prozoru (na malom prostoru)
- Prikazani sažetak sadrži ceo sadržaj prozora, a ne samo termove iz upita

Dinamički sažetak

Upit: “pretraga indeksiranje lucene”

izdvojeni fragmenti:
. . . ***Za indeksiranje i pretraživanje tekstualnih sadržaja korišćena je Apache Lucene [57] biblioteka. Apache Lucene je javno dostupna biblioteka*** pisana u Javi namenjena pretraživanju teksta. Pošto je kriterijum sličnosti definisan (u opisu slučaja korišćenja <Pick journal>) tako da su ćirilično i latinično pismo ravnopravni ***svi ćirilični sadržaji se pre indeksiranja prevode na latinično pismo, a prilikom pretrage podataka svi ćirilični upiti se prevode na latinično pismo. To znači da Apache Lucene radi samo sa sadržajima zapisanim latiničnim pismom***, ali se u bazi podataka sadržaji čuvaju onako kako ih je korisnik uneo. Prevođenje ćiriličnih sadržaja na latinično pismo je jednoznačno. . .

Google primeri za dinamičke sažetke

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .
- . . . ali moraju biti dovoljno dugački da nose neku informaciju

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .
- . . . ali moraju biti dovoljno dugački da nose neku informaciju
- Sažeci bi trebalo da opišu da li i kako dokument odgovara upitu

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .
- . . . ali moraju biti dovoljno dugački da nose neku informaciju
- Sažeci bi trebalo da opišu da li i kako dokument odgovara upitu
- Idealno: lingvistički ispravni sažeci

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .
- . . . ali moraju biti dovoljno dugački da nose neku informaciju
- Sažeci bi trebalo da opišu da li i kako dokument odgovara upitu
- Idealno: lingvistički ispravni sažeci
- Idealno: sažetak bi morao da zadovolji upit, da korisnik ne mora da pregleda dokument

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .
- . . . ali moraju biti dovoljno dugački da nose neku informaciju
- Sažeci bi trebalo da opišu da li i kako dokument odgovara upitu
- Idealno: lingvistički ispravni sažeci
- Idealno: sažetak bi morao da zadovolji upit, da korisnik ne mora da pregleda dokument
- Dinamički sažeci su važan deo zadovoljstva korisnika

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .
- . . . ali moraju biti dovoljno dugački da nose neku informaciju
- Sažeci bi trebalo da opišu da li i kako dokument odgovara upitu
- Idealno: lingvistički ispravni sažeci
- Idealno: sažetak bi morao da zadovolji upit, da korisnik ne mora da pregleda dokument
- Dinamički sažeci su važan deo zadovoljstva korisnika
  - jer možemo brzo da ih pregledamo da pronađemo relevantan dokument na koji ćemo da kliknemo

Dinamički sažetak

- Prostor na stranici sa rezultatima je ograničen → sažeci moraju biti kratki . . .
- . . . ali moraju biti dovoljno dugački da nose neku informaciju
- Sažeci bi trebalo da opišu da li i kako dokument odgovara upitu
- Idealno: lingvistički ispravni sažeci
- Idealno: sažetak bi morao da zadovolji upit, da korisnik ne mora da pregleda dokument
- Dinamički sažeci su važan deo zadovoljstva korisnika
  - jer možemo brzo da ih pregledamo da pronađemo relevantan dokument na koji ćemo da kliknemo
  - U mnogo slučajeva, uopšte ne moramo da kliknemo; time se štedi vreme

Generisanje dinamičkih sažetaka

- Odakle da dobavimo druge termove (osim onih iz upita) za sažetke?

Generisanje dinamičkih sažetaka

- Odakle da dobavimo druge termove (osim onih iz upita) za sažetke?
- Ne možemo konstruisati dinamički sažetak iz invertovanog indeksa, bar ne efikasno

Generisanje dinamičkih sažetaka

- Odakle da dobavimo druge termove (osim onih iz upita) za sažetke?
- Ne možemo konstruisati dinamički sažetak iz invertovanog indeksa, bar ne efikasno
- Moramo da keširamo dokumete

Generisanje dinamičkih sažetaka

- Odakle da dobavimo druge termove (osim onih iz upita) za sažetke?
- Ne možemo konstruisati dinamički sažetak iz invertovanog indeksa, bar ne efikasno
- Moramo da keširamo dokumete
- Pozicioni invertovani indeks kaže: term iz upita se nalazi na poziciji 4378 u dokumentu

Generisanje dinamičkih sažetaka

- Odakle da dobavimo druge termove (osim onih iz upita) za sažetke?
- Ne možemo konstruisati dinamički sažetak iz invertovanog indeksa, bar ne efikasno
- Moramo da keširamo dokumete
- Pozicioni invertovani indeks kaže: term iz upita se nalazi na poziciji 4378 u dokumentu
- Byte offset ili word offset?

Generisanje dinamičkih sažetaka

- Odakle da dobavimo druge termove (osim onih iz upita) za sažetke?
- Ne možemo konstruisati dinamički sažetak iz invertovanog indeksa, bar ne efikasno
- Moramo da keširamo dokumete
- Pozicioni invertovani indeks kaže: term iz upita se nalazi na poziciji 4378 u dokumentu
- Byte offset ili word offset?
- Keširana kopija može biti neažurna

Generisanje dinamičkih sažetaka

- Odakle da dobavimo druge termove (osim onih iz upita) za sažetke?
- Ne možemo konstruisati dinamički sažetak iz invertovanog indeksa, bar ne efikasno
- Moramo da keširamo dokumete
- Pozicioni invertovani indeks kaže: term iz upita se nalazi na poziciji 4378 u dokumentu
- Byte offset ili word offset?
- Keširana kopija može biti neažurna
- Ne keširati vrlo dugačke dokumente – samo kratak prefiks

Pojam

- Polazeći od zadatog skupa klasifikacija pokušava da utvrdi kojoj klasi ili klasama posmatrani objekat (dokument) pripada
- Pojam klasifikacije je usko vezan za pretraživanje podataka

Primena

- Klasifikacija dokumenata se koristi u više domena:
  - Istraživanje i analiza teksta i podataka
  - Procesiranje slike (utvrđivanje da li je landscape ili portrait)
  - Pretraživanje teksta (kao i drugih vrsta sadržaja)
  - itd.

Primena u pretraživanju teksta

- Koraci u pretprocesiranju
  - utvrđivanje enkodinga
  - segmentacija reči
  - utvrđivanje jezika
  - truecasing - da li reč treba da ostane napisana velikim slovima (Fed - fed, CAT - cat)
- Automatska detekcija spam strana koja ne treba da se nađu u rezultatima pretrage
- Automatska detekcija drugih vrsta strana koja ne treba da se nađu u rezultatima pretrage (npr. sexually explicit content)

Primena u pretraživanju teksta

- Sentiment detection
  - pozitivni ili negativni komentar
  - želimo da nađemo sve negativne komentare za neki proizvod, kada pročitamo negativne komentare odlučićemo da li da kupimo proizvod
- Klasifikacija e-mail-ova po folderima - spam folder
- Topic-specific ili Vertical search - pronađimi computer science strane na univerzitetima u Kini, ne mora se spominjati Kina na tim stranicama.
- Rangiranje rezultata na osnovu klasifikacije - veoma relevantan, prilično relevantan, itd.

Klasifikacija - pristupi

- Manuelna, odnosno ručna klasifikacija
  - dugotrajnost
  - gotovo nemoguće je primeniti na velike kolekcije
  - tačnost
- Klasifikacija bazirana na pravilima
  - najčešće ručno zapisanim
  - ova pravila u kontekstu klasifikacije tekstualnih sadržaja opisuju značenje pojedinih reči za smeštanje datog teksta u određenu klasu dokumenata

Klasifikacija - pristupi

- Sistem za klasifikaciju baziran sa tehnikama mašinskog učenja
  - kriterijumi za odlučivanje se utvrđuju automatski
  - sistem se obučava putem skupa obučavajućih podataka
  - sami objekti se reprezentuju skupom atributa relevantnih za postupak klasifikacije među kojima je i jedan atribut koji označava oznaku klase kojoj objekata pripada (klasni atribut)
  - skup obučavajućih podataka se obično ručno klasifikuje od strane eksperata - labeling (označavanje, anotacija)
  - nadgledani metod obučavanja
  - tokom obučavanja cilj je kreirati model (matematički) kojim se klasni atribut izražava kao funkcija vrednosti ostalih atributa
  - krajnji cilj je da formirani model omogućuje da novi objekti koji nisu bili deo obučavajućeg skupa što je moguće tačnije klasifikuju na osnovu vrednosti svojih atributa

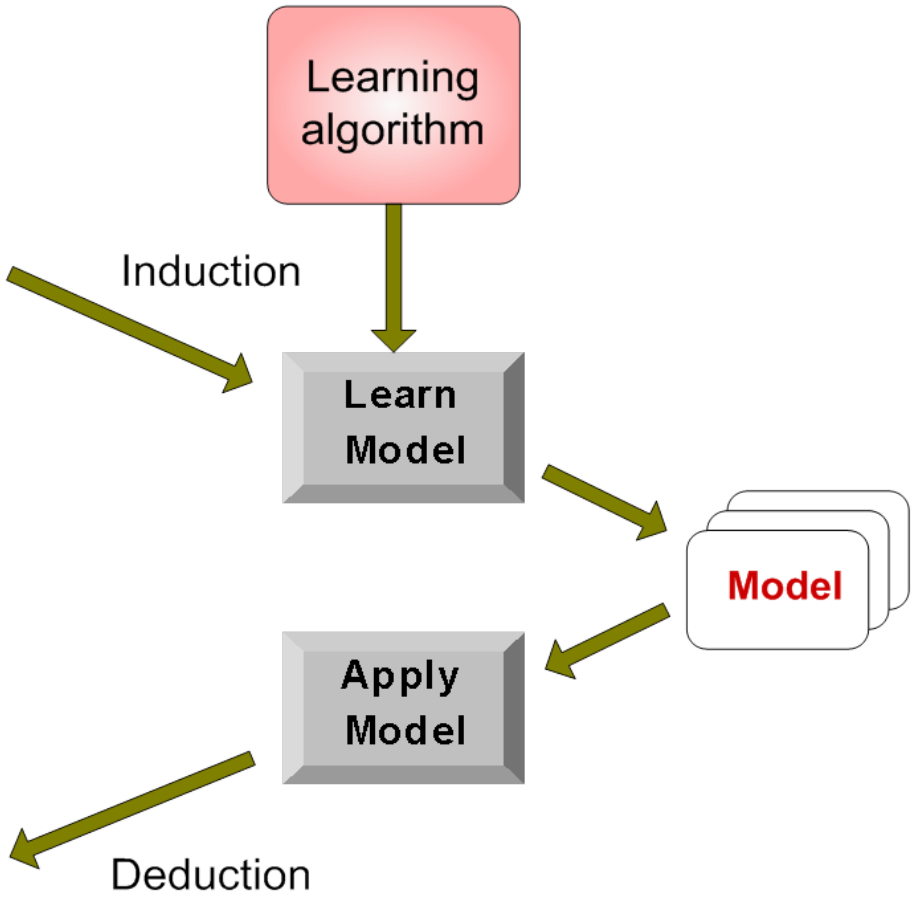
Klasifikacija na tehnici mašinskog učenja

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Metode klasifikacije

- Zasnovane na stablima učenja
- Zasnovane na pravilima
- Zasnovane na neuronskim mrežama
- Zasnovane na pravilima
- Naivni Bayes
- k najbližih suseda - kNN (k Nearest Neighbours)
- Mašine potpornih učenja - SVM (Support Vector Machines)

Kvalitet klasifikatora

- Kvalitet klasifikatora meri se na test skupu
- To je skup podataka za koji su klase poznate ali koji nije korišćen za obuku modela
- Ako nemamo takav testni skup koristimo obučavajući skup i cross-validation tehniku

Podbacivanje i prebacivanje

- Podbacivanje (underfitting) se odnosi na fenomen kod koga klasifikator na obučavajućem skupu ne daje zadate rezultate klasifikacije
- Prebacivanje (overfitting) se odnosi na fenomen kada klasifikator "slepo" sledi obučavajući skup pri klasifikaciji pa se dobijaju pogrešne klasifikacije (npr. neka reč se u obučavajućem skupu javljala samo u dokumentima koji su klasifikovani kao China, ali u realnosti ta reč nije vezana samo za Kinu)
- Sposobnost generalizacije na bazi obučavajućeg skupa
  - da se dobro klasifikuje obučavajući skup
  - da se dobro klasifikuju i novi primeri koji ne moraju da budu isti kao oni iz obučavajućeg skupa

Cross-Validation

- Ako nemamo dostupan poseban test skup, a želimo da evaluiramo klasifikator upotrebićemo obučavajući skup da iz njega izdvojimo test skup
- Osnovna ideja je deljenje svih dostupnih obeleženih podataka na dva skupa:
  - obučavajući skup – na osnovu koga se formira klasifikator
  - test skup – na kome se klasifikator evaluira
- k-tostruka unakrsna validacija
  - Podeliti skup u k jednakih delova
  - Formirai sve moguće kombinacije delova na ovaj način:
    - obučiti na (k1+...+kn-1), testirati na kn
    - obučiti na (k1+...+kn-2+kn), testirati na kn-1
    - ....
  - Izračunati prosek performansi od svih kombinacija

Metrike za evaluaciju performansi

- Tačnost
- Preciznost
- Povrat
- Matrica troškova

Pojam

- Nenadgledani metod obučavanja
- Nema definisanih klasa
- Nema obučavajućeg skupa
- Klaster analiza je podela skupa objekata na podskupove
- Cilj klaster analize je nalaženje grupe objekata takvih da su objekti iz grupe međusobno slični (as similar as possible) i da su različiti (as dissimilar as possible) od objekata iz drugih grupa

Klasifikacija vs Klasterovanje

- I jedno i drugo deli skup na podskupove
- Klasifikacija je nadgledano učenje
  - Cilj je ponovo primeniti razdvajanje klasa na novim objektima na osnovu klasifikicaje testnog skupa koju je uradio ekspert (čovek)
  - Definisane klase
  - Čovek učestvuje u inicijalnoj klasifikaciji testnog skupa

Primena

- Razumevanje
  - grupe povezanih dokumenata za pretraživanje
  - grupe gena i proteina sa sličnom strukturom
  - grupe akcija sa sličnom fluktacijom cena
  - itd.
- Sažimanje
  - smanjenje velike količine objekata

Primena u pretraživanju teksta

- Pretpostavka od koje polazimo
  - Dokumenti iz istog klastera se ponašaju **sličnо** iz ugla relevantnosti u odnosu na informacionu potrebu
  - **Verovatno** su i ostali dokumenti iz klastera relevantni
- Search result clustering
- Scatter-Gather
- Collection clustering
- Language modeling
- Cluster-based retrieval

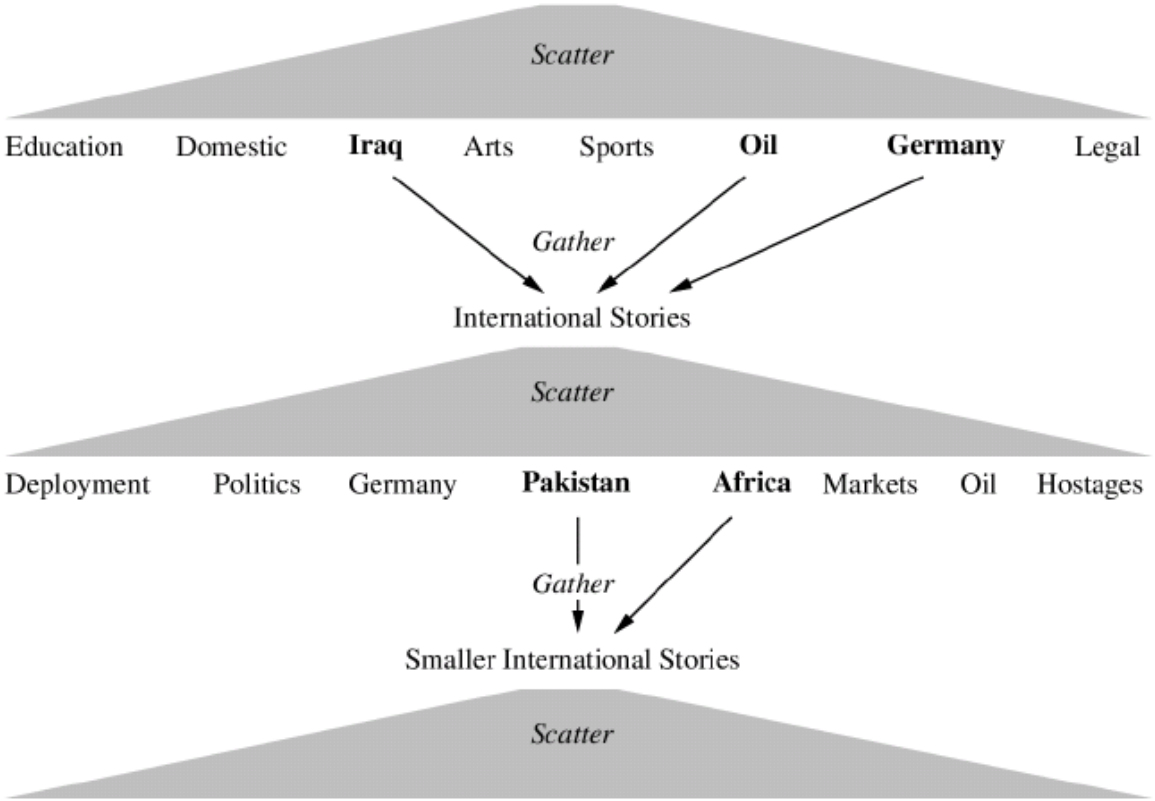
Search result clustering

- Rezultati nisu prosta lista relevantnih odgovora nego su grupisani po klasterima
- jaguar - kola, životinja, operativni sistem
- Lakše je korisniku da se snađe među odgovorima na upit

Scatter-Gather

- razbijanje na klastere i spajanje klastera
- Cilj je bolji korisnički interfejs
- Kolekcija dokumenata je klasterovana, korisnik selektuje klastere koji su mu od interesa
- Odgovori (dokumenti) koji pripadaju selektovanim klasterima su ponovo klasterovani, pa korisnik ponovo selektuje klastere koji su mu od interesa
- I sve tako dok ne dođe do malog klastera koji mu je od interesa

Scatter-Gather



Collection clustering

- Klasterovanje kolekcije, ali nema interakcije sa korisnikom
- Ako korisnik želi da čita vesti, on ih uglavnom ne pretražuje, nego želi da čita nove vesti iz neke oblasti
- Imamo veliku količinu novih vesti, korisnik želi da ispuni svoju informacionu potrebu koju retko izražava upitom
- Google News

Language modeling

- Koristimo klasterovanje da bi rešili problem sinonima
- Korisnik je u upitu koristio reč "car", pronašli smo nekoliko dokumenata koji imaju ovu reč, ali koji imaju i reči automobile, vehicle
- Vraćamo i druge rezultate iz istog klastera iako nemaju reč "car", ali imaju neku od reči automobile, vehicle zbog čega pripadaju istom klasteru
- Ovo može povećati povrat ako se uzme u obzir pretpostavka od koje smo pošli
- Ali može i smanjiti preciznost

Cluster-based retrieval

- Ubrzava pretragu
- Računanje sličnosti vektora koji predstavlja upit i dokumenata u kolekciji može biti sporo
- Alternativa poredi upit sa klasterima (kojih je znatno manje nego dokumanata) i vrati sve dokumente iz klastera
- Ovo je manje precizno od klasičnog pristupa u vektorskom modelu, ali je dosta brže, i u praksi se u nekim situacijama pokazalo da su rezultati zadovoljavajući
- Cluster pruning - iz svakog klastera se odabere par dokumenata predstavnika sa kojima se porede upiti

Algoritmi za klasterovanje

- Ravno klasterovanje
  - K-means i njegove varijacije
- hijerarhijsko klasterovanje
- klasterovanje bazirano na gustini

Validacija klastera

- Mnogo je lakše validirati nadgledanu klasifikaciju
- Sa čime da se poredimo?
- Kako da uporedimo dva algoritma za klasterovanje, koji je bolji?

Mere validnosti klastera

- Eksterni indeks
  - Meri stepen slaganja dobijenih oznaka klasa sa oznakama klasa koje su eksterno date
- Interni indeks
  - Meri kvalitet strukture klasteringa bez korišćenja eksternih informacija
- Relativni indeks
  - Poredi različita klasterovanja ili klastere, parametar u ovim indeksima mogu biti i eksterni i interni indeksi