Avtomatsko indeksiranje 1

Ročno : avtomatsko indeksiranje,

uvod v statistične metode avtomatskega indeksiranja,

krnjenje – uvod.

Reference : dokumenti

- V klasičnih tekstovnih zbirkah – bibliografskih zbirkah – je iskanje informacij v resnici iskanje referenc na informacije.
- Za izpolnitev informacijske potrebe potrebujemo dokument in ne referenco nanj.
- Bibliografska podatkovna zbirka je oddaljen približek informacijskega orodja, ki ga potrebujemo.

Razlogi za prevlado referenčnih zbirk

- Majhna zmogljivost računalnikov v času razvoja prvih bibliografskih zbirk.
- Pobudniki razvoja zbirk so bili večinoma sami raziskovalci in ne informatiki.
- Raziskovalci so poznali od informacijskih orodij le kartične kataloge.
- Bibliografska zbirka je v osnovi kartični katalog na elektronskem mediju.

Kritika deskriptorskih sistemov (1/4)

- Bibliografska zbirka kot moderniziran kartični katalog je dokončno uveljavila deskriptorski način opisovanja vsebine.
- Kritike so se osredotočale na:
  - ceno intelektualnega dela,
  - zamudnost indeksiranja.
- Kritike deskriptorskih sistemov so redko podvomile v ustreznost pristopa nasploh. Izjema je Cleverdon (1984).

Kritika deskriptorskih sistemov (2/4)

Cyril W. Cleverdon (1984):

- Če dve skupini strokovnjakov gradita tezaver za neko strokovno področje, se ujema samo 60% deskriptorjev,
- če dva izkušena indekserja indeksirata isti dokument in uporabljata isti tezaver, določita samo 30% istih deskriptorjev,
- če naredita dva informacijska posrednika poizvedbo na isto temo v isti podatkovni zbirki, je med zadetki samo 40% istih bibliografskih zapisov in
- če dva raziskovalca ocenjujeta rezultate iste poizvedbe z njunega strokovnega področja, se pri oceni relevantnosti zadetkov ujemata samo v 60%.

Kritika deskriptorskih sistemov (3/4)

- Omenjene faze predstavljajo ves postopek od priprave tezavra do iskanja po zbirki.
- Nenatančnosti v postopkih se deloma kopičijo.

Kritika deskriptorskih sistemov (4/4)

- Cleverdonova kritika je le delno utemeljena za dobro organizirana informacijska okolja s
  - strogo kontrolo gradnje in vzdrževanja tezavra,
  - standardizacijo vseh postopkov,
  - izobraževanjem informacijskih posrednikov in uporabnikov,
  - dodatnimi orodji za pomoč pri indeksiranju in iskanju.

Pomanjkljivosti klasičnega indeksiranja

- Toga pravila indeksiranja in počasnost pri uvajanju novih deskriptorjev,
- velik vložek intelektualnega dela ljudi, šolanih v stroki in s prakso v indeksiranju,
- presenetljiva ohlapnost postopkov in rezultatov pri uporabi kontroliranih tezavrov (Cleverdon, 1984).

Prednosti klasičnega indeksiranja

- Predvidljivost,
- neodvisnost od jezika dokumenta in posebnosti avtorjevega izrazja,
- enostavno avtomatiziranje širjenja in oženja poizvedbe z hierarhičnimi tezavri.

Prednosti avtomatskega indeksiranja

- Manj intelektualnega dela,
- (vsaj teoretično) reprezentirana natančno in samo vsebina dokumenta,
- (vsaj teoretično) reprezentirani vsi eksplicitno opisani aspekti vsebine dokumenta.

Pomanjkljivosti avtomatskega indeksiranja

- Ni semantičnih povezav med elementi opisa, kot jih sicer uvaja tezaver,
- velik obseg elementov vsebinskega opisa,
- jezikovna in stilistična odvisnost postopkov indeksiranja in iskanja.

Avtomatsko indeksiranje (uvod)

- Avtomatsko indeksiranje ima prednost pri zbirkah polnih dokumentov, med njimi še posebej pri
  - zelo velikih zbirkah,
  - zelo dinamičnih zbirkah.
- Teoretično bi lahko iskalni algoritmi iskali neposredno po besedilu dokumentov, iz praktičnih razlogov pa tudi avtomatsko indeksiranje zahteva predhodno obdelavo dokumenta.

Avtomatsko indeksiranje (uvod)

- Avtomatsko indeksiranje poskuša v dokumentu najti besede (ali besedne zveze), ki predstavljajo najpomembnejše vsebine.
- Take besede nosijo največjo količino informacije (povedno moč).
- Te besede (ali besedne zveze) postanejo indeksni termini.
- Pri teh postopkih ni potrebno sodelovanje informacijskega strokovnjaka.

Avtomatsko indeksiranje (uvod)

- Osnovna pristopa k obdelavi besedil pri avtomatskem indeksiranju sta
  - računalniško-jezikoslovni (lingvistični) pristop in
  - statistični pristop.

Lingvistične metode avtom. indeksiranja

- Metode poskušajo razumeti vsebino in s pomočjo razumevanja izbrati najustreznejše vsebinske predstavnike.
- Pri tem
  - uporabljajo sintaktično in semantično znanje o jeziku,
  - prepoznavajo jezikovne strukture.
- Zaenkrat metode še niso zelo učinkovite,
  - ker so računalniško potratne, in
  - formalne teorije jezika se zaenkrat ne da formalizirati v učinkovite algoritme, ki bi veljali za vse jezikovne strukture v vseh oblikah sporočanja.
- „Plitve“ metode računalniškega jezikoslovja že zelo izboljšajo kvaliteto avtomatskega indeksiranja s statističnimi metodami.

Statistične metode avtom. indeksiranja - uvod

Statistične metode temeljijo na enostavni frekvenčni analizi besedil.

- Osnovne predpostavke:
  - besede niso slučajno porazdeljene po besedilih,
  - frekvenca pojavljanja neke besede je pozitivno povezana s pomembnostjo vsebine, ki jo ta beseda zastopa,
  - besede, ki se v besedilu pojavljajo večkrat, v splošnem več prispevajo k njegovi vsebini.

Statistične metode avtom. indeksiranja - uvod

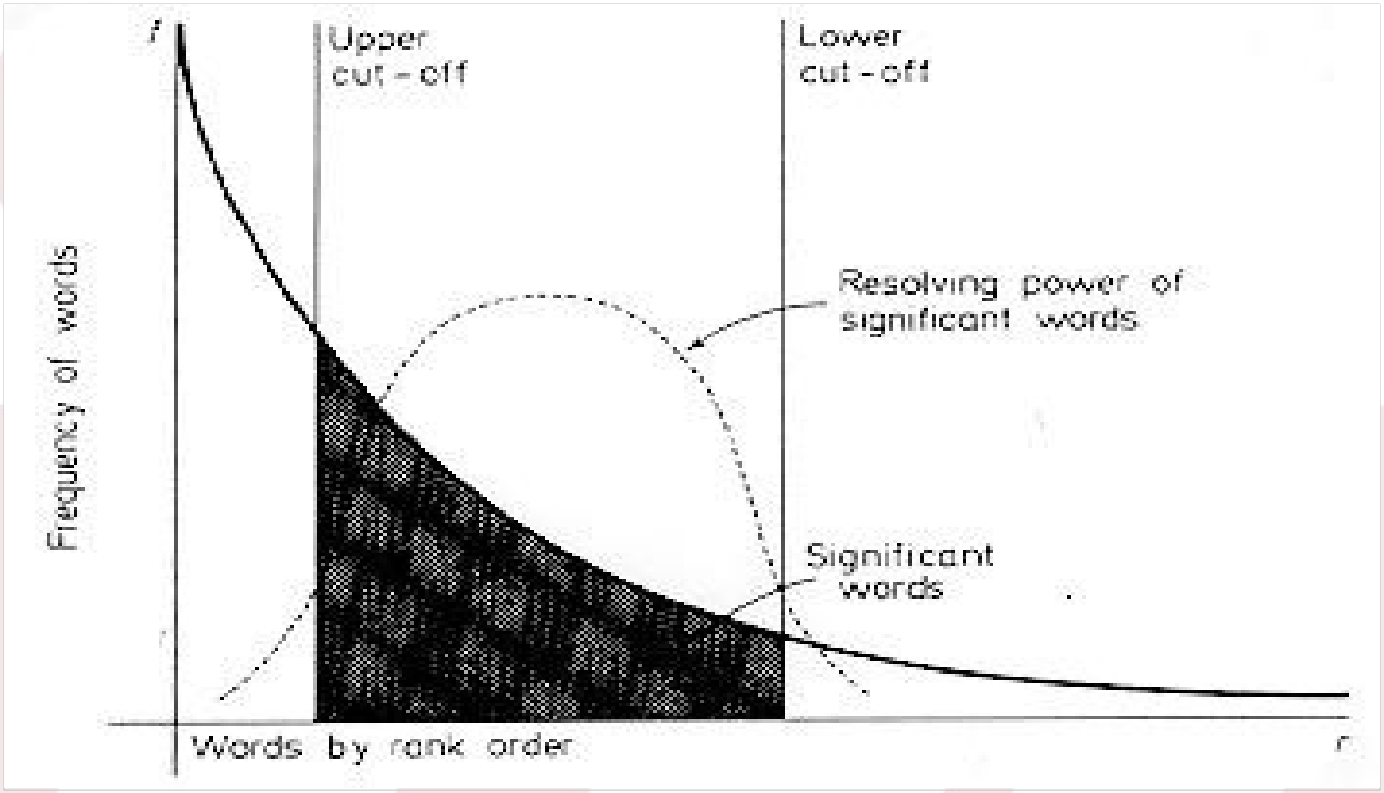
Besede iz dokumenta je treba preoblikovati tako, da so primerne za vlogo vsebinskih predstavnikov – indeksnih terminov.

- Običajno zaporedje postopkov je:
  - blokiranje,
  - krnjenje,
  - računanje povednih moči.
- Veliki spletni iskalniki krnjenje izvajajo manj radikalno, kot bi ga lahko (kot že dolgo znamo).

Blokiranje

- Izvor ideje v delu Luhna in Zipfa.
- Besede iz korpusa angleškega jezika sta preštela in razvrstila po rangih frekvenc.
- Odvisnost med frekvenco besede in njeno pozicijo v rangu je hiperbolična funkcija.

Blokiranje



- Besede v korpusu lahko navidezno razdelimo na tri skupine:
- zelo pogoste, ki ne “nosijo” vsebine dokumenta,
- zelo redke, ki niso primerne za vsebinske predstavnike, in
- tiste “vmes”, ki “nosijo” vsebino.

Blokiranje

- V skupini zelo pogostih besed je malo različnih besed. Te besede se pojavljajo v vseh besedilih v nekem jeziku.
- Ker nosijo malo informacije, jih lahko zavržemo.
- Te besede sestavljajo t.i. seznam blokiranih besed (stop-words list).
- Blokirane besede sodijo v nekatere besedne vrste: predlogi, prislovi, zaimki...

Blokiranje

- V angleških besedilih je število zelo pogostih različnih besed manjše kot pri slovenščini, ker za eno angleško besedo obstaja veliko besednih oblik slovenskega prevoda.
- Seznami blokiranih besed za slovenščino so zato precej večji od angleških.

Blokiranje

avtor	**št. blokiranih besed**
van Rijsbergen (angl.)	250
Brown corpus (angl.)	425
Popovič (slov.)	1593
Dimec (slov.)	1205

Različne velikosti seznamov blokiranih besed za slovenščino in angleščino

Krnjenje

- Postopek, s katerim nevtraliziramo morfološko bogastvo jezika.
- Besede v dokumentih nastopajo v različnih pojavnih oblikah (zaradi sklanjanja, spreganja, števila, spola…).
- Pri krnjenju (stemming) iščemo zaporedje znakov, ki lahko zastopa vse oblike neke besede in samo oblike te besede.

Krnjenje

- Krnjenje ne določa pravega korena besede, zato raje govorimo o krnih.
- Koren in krn sta pogosto enaka.
- Krn ni nujno vsebovan v vseh oblikah neke besede
- S krnjenjem dosežemo isto kot z “ročnim krajšanjem” pri oblikovanju iskalne zahteve, le da ga opravimo že pri vključevanju dokumenta v zbirko.

Krnjenje

Osnova so predpostavke:

- besede z dovolj dolgim enakim zaporedjem začetnih znakov so tudi vsebinsko sorodne (dokaz: etimologija besed);
- pojavljanje različnih končnih delov besed s skupnim začetnim zaporedjem se ravna po nekih pravilih (dokaz: morfološka pravila);
- ta pravila so dovolj enostavna, da jih je mogoče formalizirati v ekonomičen algoritem.

Krnjenje

Obstajata dve široki skupini algoritmov:

- algoritmi brez seznama kočnic
  iščejo skupne krne z upoštevanjem nekaterih statističnih posebnosti besed,
- algoritmi s seznamom končnic
  oblikujejo krne tako, da od besed režejo njihove končnice.