VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS KATEDRA

Baigiamasis bakalauro darbas

Dokomentų klasterizacija

(Document clustering)

Atliko: 4 kurso 1 grupės studentas
Dominykas Ablingis (parašas)
Darbo vadovas:
lekt. Rimantas Kybartas (parašas)
Recenzentas:
doc. dr. Vardenis Pavardenis (parašas)

Vilnius
2018

Turinys

Sąvokų apibrėžimai

- cluster
- clustering
- document
- corpus
- term – termas dar vadinamas (tokens, words,terms or attributes)
- token
- feature
- online
- ofline
- supervised
- unsupervised
- matmenų redukcija (angl. *dimensionality reduction*)

Įvadas

Šiais laikais kai kiekvienas žmogus turintis priėjimą prie interneto gali dalintis informacija, daugybė knygu yra skaitmenizuojamos kiekvieną dieną ir mokslo institucijos dalinasi savo moksline informacija, pasiekemos informacijos kiekis didėja su kiekviena diena ir pagrindinė problema nebėra informacijos trūkumas, o atradimas ko reikia. Tam spresti buvo ir yra kuriama ivairūs mechanizmai: paiškos varikliai...Šiame darbe autorius nagrinės viena iš šios problemos sprendimo metodų, klasterizacijos algoritmus ir jų panaudojimą textinems dokumentams.

Šiandieninėje visuomenėje prieinamos informacijos kiekis didėja su kiekviena diena ir pagrindinė problema yra ne informacijos trūkumas, o jos gausa ir galimybė surasti reikiamą informaciją. Paieškos programos gali pateikti didelius kiekius tekstinių dokumentų, bet reikiamo dokumento paieška užima daug laiko ir be to, ne visada gauta informacija atitinka paiešką. Tekstinių dokumentų paieškos procesui palengvinti ir pagreitinti gali būti taikomi klasterizavimo metodai, kurie yra pakankamai gerai žinomi ir jau seniai naudojami duomenų klasterizavimui. Klasterizavimo metu dokumentai pagal savo turinį suskirstomi į klasterius vadovaujantis tam tikrais kriterijais, pvz., pagal temą, pagal dokumento naujumą ir pan.

1. Pagrindinė tiriamoji dalis

Darbo tikslas – mokslinės literatūros apie dokumentų apdorojimą ir klasterizavimą apžvalga ir analizė

Darbo uždaviniai:

1. Panaudojimo sritys
2. Dokumentų klasterizavimui naudojamų įrankių ir metodų apžvalga.
3. Dokumentų klasterizavimo proceso žingsnių aprašymas
4. Kylantys iššūkiai
5. Susijusios problemos

Kursiniame darbe buvo siekta susipažinti su moksline literatūra, išanalizuoti egzistuojančius dokumentų klasterizavimo metodus ir įrankius, kurie bus išbandyti taikomojo pobūdžio kursiniame darbe apie lietuviškų dokumentų klasterizavimą.

Šiame darbe taip pat bus paminėtos kitos su dokumentu klasterizacija susijusios problemos ir ši informacija bus išdėstyta eilės tvarka kaip būtu sprendžiamos užduotys.

1. Duomenų išgavimas iš skirtingų tekstinių dokumentų formatų
2. Duomenų apdorojimas
3. Algoritmai
4. Algoritmų testavimas
5. Rezultatų vizuoalizacija

2. Algoritmų testavimas/kokybės vertinimas

viena iš fundamentalių neprižiūrimo mokymosi problemų yra modelių testavimas. skirtingai nei „prižurimame mokymasi“ kur svarbu atdidėti dali duomenų su kuriais nėra mokomasi o tik testuojama sugeneruoti modeliai, „neprižiurimame mokymasi“ mes to nelgalim atlikti...bet egzsituoja keltatas metodu kaip galima patikrinti sudarytus klasterius.

- pirma galima sugeneruotus klasterius leisti tikrinti **žmonėms** . tam yra keli būdai. galima tiesiog duoti sugeneruotus klasterius ir bandyti nuspresti ar jie tinkami. taipat galima parainkti du atsitiktinius dokumentus ir spėti ar jie turėtu būti viename ar atskiruose klasteriuose, ir tada palyginti su kompiuterio sugeneruotu rezultatu.

- taip yra metodų kaip atlikti testavimą automatiškai. vienas jų paimti 2(ar daugiau) dokumentus iš skirtingų klasterių ir apkeisti juos vietomis, tada patikrinti klasteriu //stipruma. tai atlike dokybe kartų galime spręsti kaip sėkmingai sekėsi klasterizacija, jeigu apmainant jų kokybė nukrito tai reikškia, kad dokumentai sėkmingai suklaserizuoti, bet jeigu nesikeite tai reiške, kad klasteriai mažai vienas nuo kito skiresi ir klasreizacija nesekminga.

3. Rezultatų vizuoalizacija

dažnai norėdami geriau pažinti (ar testuoti) klasterizacijos rezultatus mes galime juo vizuolizuoti. vizulaizacijos gali buti įvairios ir dažnai priklauso nuo algoritmo rūšies, bet dažniausiai naudojama **point cloud** vizualizacijos. jose matosi pagal kokius parameturs (ašis) buvo klasterizuojama ir kaip atrodo sudaryti klasteriai.taip pat galima pridėti papildomų indikatorių, priklausomų nuo algoritmo. pavyzdžiui klasteriams sudarytiems k-means metodu galima nubraižyti atitinkamus „centrinius taškusn".

4. Susijusios, aktuoalios problemos

4.1. keyword extraction

4.2. cluster labeling

4.3. document classification

4.4. dimensionality reduction

dimensionality reduction methods can be considered a subtype of soft clustering; for documents, these include latent semantic indexing (truncated singular value decomposition on term histograms) and topic models.

Išvados

Conclusions

Šiame skyriuje pateikiamos išvados (reziume) anglų kalba.