Vadyba
Journal of Management
2009, Nr. 3(16)
ISSN 1648-7974

ŽYMĖJIMU IR ONTOLOGIJOS TAIKYMAS GRINDŽIAMAS TEKSTINIŲ DOKUMENTŲ KLASTERIZAVIMU

Marijus Bernotas, Asta Slotkienė
Vakarų Lietuvos verslo kolegija

Anotacija
Šiuolaikinėse organizacijose tekstiniai dokumentai atlieka labai svarbų vaidmenį, o nuolatinis jų kaupimas didina dokumentų saugyklų apimtis. Dažniausiai siūlomi informacijos gavybos iš tekstinių dokumentų metodai yra pagrįsti tik žodžių atitikimo tikrinimu. Alternatyvus informacijos išrinkimo būdas – klasterizavimas. Pastaraisiais metais aktyviai vykdomi įvairūs tyrimai, kurių metu siūlomos vis naujos klasterizavimo metodų modifikacijos, panaudojamos žinių inžinerijos technologijos. Šiame straipsnyje siūlome papildyti tradicinį klasterizavimo metodą žymėjimu grindžiama dokumentų atvaizdavimu, o klasterizavimo rezultatus atvaizduoti panaudojant žinių technologiją – ontologiją
PAGRINDINIAI ŽODŽIAI: tekstinių dokumentų klasterizavimas, ontologija.

Įvadas

Informacijai išrinkti praktikoje naudojami įvairūs paieškos metodai, kurie šiuo metu yra nuo paprasčiausios paieškos pagal žodį iki intelektualios paieškos pagal semantinę koncepciją. Tačiau paieškos taikymas duomenų gavybai reikiamą rezultatą suteikia tik po kelių ieškojimų ciklų. Tai mažina informacijos išrinkimo efektyvumą. Dokumentų klasterizavimas siūlo alternatyvų būdą greitam reikalingos informacijos suradimui. Skirtingai nei vykdant paieškos užklausas, klasterizuojant dokumentai į bendras grupes – skirstinius (angl. *cluster*) – patenka ne pagal ieškomų žodžių ar frazių buvimą ar nebuvimą juose, bet pagal dokumentų savybių (tekstinio dokumento turinį) panašumą. Pastaruoju metu šis būdas sparčiai tobulinamas ir siūlomos įvairios modifikacijos klasterizavimo metodų lygmenyje ir duomenų gavybos metodų derinimo aspektu.

Nepriklausomai nuo pasirenkamo klasterizavimo metodo, jo metu atliekami šie žingsniai (Theodoridis; Koutroumbas, 2003): dokumentų atvaizdavimo parinkimas, panašumo mato parinkimas, klasterizavimo metodo parinkimas, grupių pateikimo pasirinkimas, rezultatų validavimas. Įprastus klasterizavimo metodus taikant tekstiniams dokumentams, kurie parengti mažai paplitusiomis kalbomis, tokios kaip, lietuvių, latvių, estų, švedų, olandų, ir t.t., pastebėta, kad sudaromas nekompaktiškas žodžių ir frazių žodynas, kuris naudojamas dokumentų atvaizdavimui. Šiai problemai spręsti siūlome taikyti žymėjimo technologiją, o tinklinę dokumentų struktūrą atvaizduoti panaudojant ontologiją.

Dokumentų klasterizavimas ir klasifikavimas

Tekstinių dokumentų klasifikavimas įvairiose organizacijose yra plačiai paplitęs ir daug kur taikomas dėl savo technologinio įgyvendinimo paprastumo. Tačiau klasifikuodami dokumentus dirbama su iš anksto apibrėžta kategorijų aibe ir klasifikacijos proceso metu priklausomai nuo turinio, dokumentas yra priskiriamas kuriai nors vienai iš kategorijų, o klasterizuojant dokumentus jokios iš anksto apibrėžtos kategorijų aibės nėra. Dokumentai yra suburiami į grupes tik pagal savo savybių panašumą t.y., dokumentų grupės yra suformuojamos kiekvieną kartą pagal įvairias dokumentų savybes, o ne klasifikuojami į iš anksto suplanuotą organizacijos kategorijų sistemą.

Aplinkos objektų klasterizavimas yra įgimta kiekvieno žmogaus savybė. Įprasta aplinkinius asmenis intuityviai skirstyti į vyrus ir moteris, daiktus skirstyti pagal geometrines figūras, paskirtį, spalvas ir pan. Jau prieš 30 metų dokumentų klasterizavimas buvo taikomas informacijos išrikimui. Mokslininkai kasmet siūlo vis naujų klasterizavimo metodų modifikacijų, kurios gali atsižvelgti ne tik į objektų tarpusavio panašumus, bet ir santykių tarp jų pobūdį. To pasekoje gaunami tikslesni paieškos rezultatai, kurie labiau atitinka vartotojų užklausas. Literatūroje yra aprašyta daug klasterizavimo metodų, kurie tarpusavyje skiriasi dokumentų atvaizdavimo aprašymu, taikomais panašumo matais ir klasterizavimo algoritmu bei dokumentų grupių pateikimo būdu. Šiame darbe plačiau panagrinėsime dokumentų atvaizdavimo būdus ir jų tobulinimo tendencijas.

Dokumentų atvaizdavimo būdai

Vienas iš pirmų žingsniu, įtakojančių klasterizavimo rezultatą, yra dokumentų atvaizdavimo

būdo parinkimas. Tai atributų parinkimas, kurie, taikant klasterizavimo algoritmą, atvaizduoja kiekvieną dokumentą. Dažniausiai dokumento atvaizdavimui yra panaudojamas dokumento tekstas. Praktikoje tekstinių dokumentų klasterizavimui plačiausiai taikomas vektorinės erdvės modelis. Šiame modelyje dokumentas apibrėžiamas euklidinės erdvės vektoriumi, kuriame tekstiniai dokumentai apibūdinami žodžių vektoriais $<t_1, ... t_n>$, kur t_i yra žodžio svoris dokumentų kolekcijoje, o n yra klasterizuojamos dokumentų kolekcijos žodyne esantis žodžių skaičius. Kolekcijos žodynas yra sudaromas iš dokumento tekste esančių žodžių, atliekant žodžių filtravimo procesą, kurio metu atmetamos skirtingos formos žodžiai, semantiniai atitikmenys ir tokios kalbos dalis kaip įvardžiai, prielinksniai ir pan. Pastebėta, kad vektorinės erdvės modeliui būdingi šie ypatumai:

1. Nesudėtingas taikymas dokumentų savybių aprašymui;
2. Paprastai apskaičiuojamas panašumas tarp dviejų dokumentų;
3. Klasterizavimo rezultatą įtakoja kolekcijos žodynas, kuriame gali būti ignoruojami kelių žodžių išreiškimai, kaip, pavyzdžiui, Europos sąjunga. Taip pat problematiškas sinonimų ir daugiareikšmių sinonimų interpretavimas, kadangi jiems gali būti priskirtos tapačios savybės
4. Neoptimalus taikymas lokaliai vartojamoms kalboms, tokioms kaip olandų, latvių, lietuvių ir pan., kadangi sudaromas nekompaktiškas žodynas ir gali būti iškreipiami klasterizavimo rezultatai.
5. Dokumentų savybių išskyrimui ignoruojamas žodžių apibendrinimo ryšys. Kolekcijos žodyne žodžiai pateikiami nepriklausomai vienas nuo kito, tačiau dažniausiai jie priklauso tam tikrai grupei, kaip pavyzdžiui, auksas ir sidarbas priklauso brangiesiems metalams.

Minėtų trūkumus (3-5) siūlome spręsti taikant sparčiai populiarėjančią žymių (angl. *tags*) priskyrimo objektams technologija, kuri vadinama žymėjimu (angl. *tagging*). Žymės - tai yra tekstinių dokumentų turinius aprašantys metaduomenys, kurie dar vadinami raktažodžiai (angl. *keywords*). Žymėjimo technologijos esminis privalumas, kuris naudingas klasterizavimo procese, yra objekto priskyrimo kelioms kategorijoms galimybė. Klasifikavimo mechanizmas, grindžiamas objektų žymėjimu, taikomas įvairių skaitmeninių objektų aprašymui. Dažniausiai tai atlieka dalykinės srities specialistai. Scott A. Golder ir kt. autorių atliktame tyrime pastebėta, kad suteikiant galimybes priskirti objektams žymes ne specialistams, o jų autoriams ar kitiems objekto vartotojams sudarant galimybes priskirti savo ir validuoti kitas žymes, sukuriama stebėtinai stabili ir dėsninga klasifikacijos sistema.

Remiantis žymėjimo technologija, klasterizuojami dokumentai siūlomame metode yra aprašomi dokumentų autorių priskirtomis žymėmis. Tuomet modifikuotame vektorinės erdvės modelyje dokumentai aprašomi žymėmis, kur dokumentas prilyginamas daugiamatės erdvės vektoriui žymių vektoriais $<t_1, ...t_m>$, kur t_j žymės svoris kolekcijos dokumente, o m yra klasterizuojamos dokumentų kolekcijos žodyne esantis žymių skaičius. Kadangi žymes priskiria dokumentų autoriai ir jų kiekis gali būti skirtingas dokumentų kolekcijoje, tai t_j jų svoriai papildomai normalizuojami taip, kad dokumentų vektorių ilgis būtų lygus vienetui. Tam taikome šią formulę:

$$t_j = \frac{a_j \times df_j}{\sqrt{\sum_{r=1}^{m} (a \times df_r)^2}} \qquad (1)$$

čia a_j (a_r) nurodo ar j-oji (r-oji) dokumentų kolekcijos žymių žodyne esanti žymė priskirta dokumentui.

Remiantis tyrimų duomenimis (Golder, Huberman, 2006) galime teigti, kad subjektyvus ir individualus žymių priskyrimas dokumentui jo autoriaus, neturi įtakos nepilnaverčiam kolekcijos žodynui, kadangi paparastai vartotojai parenka bendrus terminus, apibūdinančius dokumento turinį.

Tyrimo rezultatai

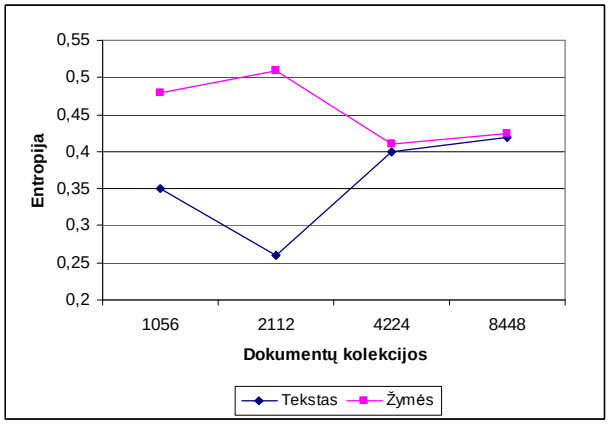
Eksperimento metu buvo sudarytos keturios skirtingo dydžio tekstinių dokumentų kolekcijos iš 13769 dokumentų archyvo (1 lentelė). Tyrimo metu klasterizuojamos dokumentų kolekcijos panaudojant skirtingus atvaizdavimo būdus: dokumentus atvaizduojant dokumento tekstu arba atvaizduojami jam priskirtomis dokumentų autorių žymėmis. Vidutinis priskiriamų žymių kiekis dokumentui svyruoja nuo 1 iki 7 žymėmis. Eksperimento metu sudarytos 5 dokumentų klasės.

1 lentelė. Dokumentų kolekcijos

Dokumentų kolekcija	Dokumentų kiekis	Žodžių kiekis kolekcijoje	Žymių kiekis kolekcijoje
1	1056	36843	1379
2	2112	55352	1882
3	4224	81293	2449
4	8448	116755	2754

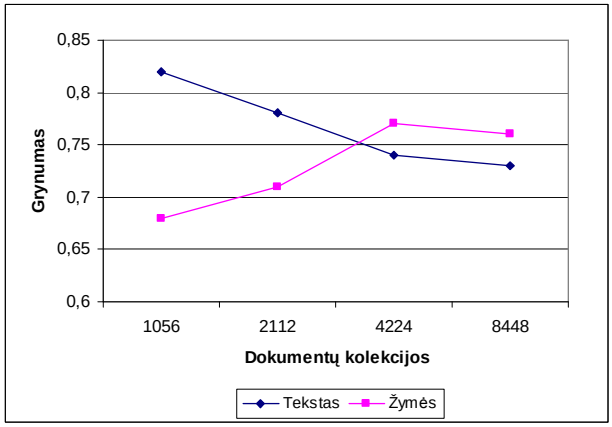
Klasterizavimo procese dokumentų panašumui įvertinti taikomas kosinuso koeficientas, naudojamas skaldantis į 5 skirstinius (dokumentų grupes) K-*means* algoritmas. Gauti rezultatai įvertinami dvejomis metrikomis: entropija, grynumas.

Entropija, kuri nusako skirtingų klasių dokumentų išsisklaidymo laipsnį vieno skirstinio ribose. Kuo entropijos reikšmė artimesnė 0, tuo išsisklaidymo laipsnis mažesnis ir klasterizavimo rezulatai geresni.



1 pav. Entropijos koeficientas įvairiose dokumentų kolekcijose

Iš entropijos grafiko (2 pav.) matyti, kad entropijos įvertinimas geresnis, dokumentų klasterizavimui taikant jo tekstą, o ne autorių priskirtas žymes. Tai galėjo įtakoti nedidelis žymių kiekis (nuo 1 iki 7) priskiriamas dokumentui. Tačiau, didėjant dokumentų skaičiui (4224 ir 8447 dokumentų kolekcijose), entropijos vertės tampa panašios.



2 pav. Grynumo vertės įvairiose dokumentų kolekcijose

Grynumas įvertina vienos klasės dokumentų susitelkimo laipsnį vieno skirstinio ribose. Grynumo vertė (2 pav.), esant didesnei tekstinių dokumentų kolekcijai, pranašumas pastebimas kuomet naudojamas žymėjimu grindžiamas klasterizavimo metodas, o ne iš tekstinio dokumento sudarytų žodžių rinkiniais.

Apibendrinus tyrimo rezultatus, galime teigti, kad naudojant nekompaktiška žodyną (žodžiai iš dokumento teksto mažai paplitusioje lietuvių kalboje) grindžiama klasterizavimo metodą, gaunami kokybiški klasterizavimo rezultatai, tačiau, esant dideliam dokumentų skaičiui kolekcijoje, žymėjimu grindžiamas metodas net lenkia grynumo metrikos rezultatus.

Ontologijos panaudojimas dokumentų atvaizdavimui

Iki 1990 metų žodis ontologija buvo vartojamas tik filosofijos mokslų srityje. Ir tik vis labiau didėjant informacijos svarbai, šis terminas pradėtas vartoti žinių inžinerijoje. Kompiuterijos mokslas ontologija apibrėžia, kaip žinojimo sritį, kuri nusako terminų, naudojamų tam tikrame tekste reikšmę arba reikšmių sritį, jų tarpusavio sąryšius ir priklausomybę (Čaplinskas A. ir kt.). Informacijos išrinkimui iš tekstinių dokumentų ontologija gali būti naudojama kaip dokumentų kolekcijos žodyno konceptuali specifikacija (Cheng K ir kt.), kuri apibrėžia ne tik žodžius ar jų junginius, bet jų semantinius ryšius: semantinis panašumas, tarpusavio priklausomybės apibrėžtis, apibendrinimo ryšius.

Pastebėta, kad dokumentų reprezentavimo metodas, grindžiamas vektorinės erdvės modeliu, klasterizavimo rezultatams turi šiuos neigiamus padarinius:

1. Kelių priklausomų žodžių išreiškimo pilnumo stoka. Pavyzdžiui, „Europos sąjunga“, žodis Europos būtų ignoruojamas.
2. Sinonimams priskirtos skirtingos savybės. Pavyzdžiui, agentas, atstovas, tarpininkas. Pagal prasmę jie sinonimiški, tačiau priklausomai nuo konteksto interpretuojant gali būti priskirtos netapačios savybės.
3. Daugiareikšmiai žodžiai gali būti interpretuojami pagal vieną savybę, o suprantami kontekste skirtingai. Pavyzdžiui, žodis virusas. Priklausomai nuo konteksto gali reikšti virusą plintantį kompiuterių sistemose arba žmogaus ligos sukėlėją.
4. Sąvokų apibendrinimo nebuvimas. Sąvokos yra nepriklausomos. Pavyzdžiui, auksas ir sidabras priklauso brangiesiems metalams.

Pirmieji trys trūkumai išsprendžiami dokumentų kolekcijos žodyno sudarymo lygmenyje, o paskutinysis reikalauja konceptualaus sprendimo. Holto A. ir kt. autoriai siūlo papildyti vektorinės erdvės modelį panaudojant ontologiją, kurį leistų atsižvelgti į ryšius tarp žodžių dokumentų kolekcijoje neprarandant dokumento konteksto. Jie siūlo vektorinės erdvės modelyje integruoti ontologiją. Taikant ontologiją, dokumentas būtų aprašytas daugiamatės erdvės vektoriumi, kuris apimtų ne tik dokumente esančius žodžius (tf nusako žodžio t_i pasikartojimų dažnį dokumente d), bet ir konceptus (cf yra koncepto c_j pasikartojimų dažnis dokumente) joje:

$$t_d = < tf(d,t_1)...tf(d,t_m), cf(d,c_1)...cf(d,c_l) \qquad (2)$$

Taikant ontologijomis grindžiama klasterizavimo metodą, kaupiama informaciją ne tik apie dokumente esančius žodžius ar frazes, bet ir apie ryšius tarp jų. Konstruojamas modelis sudarytas iš ontologijų ir dokumentus aprašančių žodžių vektorių. Atlikti tyrimai leidžia teigti, kad žinių panaudojimas dokumentų atvaizdavimui gerina grynumo įverčius. Ypatingai aktualu panaudoti siūlomą metodą ir žymėjimu grindžiamame klasterizavimo metode, kadangi dokumento autoriai priskiria ribotą kiekį (dažniausiai neviršija 10 žymių) raktažodžių.

Tradicinėse dokumentų klasterizavimo sistemose tekstiniai dokumentai kaupiami jų saugykloje, o vykdant jų išrinkimą klasterizavimo metodu išrenkami panašūs dokumentai. Naudojant ontologiją

dokumentų skirstinių sudarymas tampa efektyvesnis, kadangi yra naudojama ir ontologijos saugykla, kurioje kaupiami sąryšiai tarp dokumentų jų kolekcijoje apibūdinančių sąvokų.

Kadangi ontologijos panaudojimas vektorinės erdvės modelyje suteikia sąvokų semantinius ryšius ir apibendrinančius, tai galime teigti, kad jos leidžia sudaryti aukšto lygmens struktūrinį vaizdą visoje dokumentų kolekcijoje ar dokumentų grupėje ir sudaro galimybę atvaizduoti nestruktūrinius dokumentus dokumentų saugyklose.

Išvados

Informacijos išrinkimas iš hierarchiškai saugomų dokumentų jų saugyklose neoptimalus, kadangi dokumentas priklauso tik vienai kategorijai. Archyvuose gausu dokumentų, kurie pagal savo turinį priklauso daugiau nei vienai kategorijai. Taikant viena iš informacijos išrinkimo būdu – klasterizavimą, minėtas trūkumas pašalinamas. Eksperimento metu klasterizavimo metodas taikytas dokumentų kolekcijoms, kurios skyrėsi dokumentų atvaizdavimo būdu: žodžiai iš dokumento teksto, dokumento autoriaus priskirtos žymės. Tyrimo rezultatai leidžia teigti, kad žymėjimu grindžiamas klasterizavimo metodas rezultatus veikia neigiamai, esant negausiai dokumentų kolekcijai, ir pateikia kokybiškus rezultatus, kai dokumentų kolekcijoje daug. Pagrindinė to priežastis – nedidelis žymių kiekis, priskirtas dokumento atvaizdavimui. Remiantis pastarųjų metų atliktais tyrimais, galime teigti kad šis trūkumas ištaisomas atsižvelgiant ne tik į tiesiogiai du dokumentus vienijančias savybes, bet ir jų kontekstą. Tai įgalina ontologijos panaudojimas dokumentų atvaizdavimo etape, kadangi būtų kaupiami ryšiai tarp žymių dokumentų skirstiniuose ir dokumentų kolekcijų ribose, o tada atsiranda galimybė panaudoti sąvokų apibendrinimo ryšį. Manome, kad jos integravimas leistų pagerinti ir dokumentų panašumo įverčius

Literatūra

Cheng, K. Ch., Pan, X., Kurfess, F. *Ontology-based Semantic Classification of Unstructured Documents*. (2003). [Žiūrėta 2009 lapkričio 5d.], <http://eil.stanford.edu/ xpan/Cheng-Pan-Kurfess-2003.pdf>.

Bloehdorn, S., Cimiano, P., Hotho, A., Staab S. An Ontology-based Framework for Text Mining. (2005). *LDV Forum – GLDV Journal for computational linguistics and language technology*. 20(1), 87-112

Čaplinskas, A., Čiukšys, D. (2003). Ontologijų panaudojimo ypatumai kuriant moderniąsias informacines sistemas. *Informacijos mokslai*, 26, 94-97.

Hotho, A. Maedche, S. Staab, V. Zacharias. (2002). *On Knowledgeable Supervised Text Mining to appear in: "Text Mining" Workshop Proceedings*, Springer.

Hotho, S. Staab, G. Stumme. Text Clustering Based on Background Knowledge. *Technical Report 425*, University of Karlsruhe, Institute AIFB, 76128 Karlsruhe, Germany. April 2003

Golder, S., Huberman, B. A. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 198–208.

Weiss, D. (2006). Descriptive Clustering as a Method for Exploring Text Collections. PhD thesis. Poznan University of Technology, Poznań.

Chang, H-C. Hsu, C-C. (2005). *Using topic keyword clusters for automatic document clustering*, 419- 424.

Bernotas, M., Laurutis, R., Slotkienė, A. (2005). Issues on Forming Metadata of Editorial System's Document Management. *Information Technology And Control*, 34(4), 371 - 376.

Bloehdorn, S., Cimiano, P., Hotho, A., Staab, S. An Ontology-based framework for text mining, LDV Forum – GLDV. *Journal for computational linguistics and language technology*, 20(1), 87-112.

Theodoridis, S., Koutroumbas, K. (2003). *Pattern Recognition*, Second Edition. Academic Press, San Diego.

TEXT DOCUMENT CLUSTERING USING TAGGING-BASED TECHNIQUE AND ONTOLOGY

S u m m a r y

Text documents are very significant in contemporary organizations, and their constant accumulation enlarges the scope of document storage. Standard text mining and information retrieval techniques of text document usually rely on word matching. An alternative way of information retrieval is clustering. In these latter years various researches are performed; new modifications of clustering methods are proposed; knowledge engineering technologies are used. In this paper we suggest to complement the traditional clustering method by document representation based on tagging, and to improve clustering results by using knowledge technology – ontology.

KEYWORDS: text document clustering, ontology.

Marijus Bernotas, VšĮ Vakarų Lietuvos verslo kolegijos Šiaulių fakulteto lektorius. El. paštas marijus.bernotas@tf.su.lt.
Asta Slotkienė, VšĮ Vakarų Lietuvos verslo kolegijos Šiaulių fakulteto lektorė. El. paštas slotkiene@tf.su.lt.