doi:10.3359/oz1201006

članek

KLASIFIKACIJA KLJUČNIH BESED IZ BIBLIOGRAFSKIH ENOT

Andrej Korošec

Institut informacijskih znanosti Maribor

Kontaktni naslov:
andrej.korosec@izum.si

Izvleček

Iz bibliografskih podatkov slovenskih raziskovalcev v sistemu COBISS ni preprosto določiti raziskovalnega področja, ki je obravnavano v bibliografskem viru. Podatek, iz katerega je mogoče razbrati raziskovalno področje, so ključne besede. Za cilj smo si postavili razvrščanje (klasificiranje) ključnih besed iz bibliografskih enot raziskovalcev. Za izhodišče smo izbrali klasifikacijo ARRS. Da bi dosegli najboljše rezultate, smo posebno pozornost namenili izbiri in pripravi besedil za strojno učenje. Za razvrščanje ključnih besed iz bibliografskih enot raziskovalcev smo uporabili orodje Oracle data miner za podatkovno rudarjenje. Izbrali smo metodo podpornih vektorjev, ki se uspešno uporablja pri klasifikaciji besedil.

Ključne besede

podatkovno rudarjenje, rudarjenje besedil, metoda podpornih vektorjev (SVM)

Abstract

A research field dealt with in a bibliographic source is not easily identifiable from the COBISS bibliographic data on Slovenian researchers. Keywords are the best way to identify research fields. Our aim was to sort (classify) keywords of researchers' bibliographic units. To this end, we have chosen as the classifier the classification of the Slovenian Research Agency. A special attention has been devoted to the selection and preparation of texts intended for machine learning in order to achieve best results. To classify keywords from the researchers' bibliographical units, we have used Oracle Data Miner and the support vector machines method which has already proven to be useful for the classification of texts.

Keywords

data mining, text mining, support vector machine method (SVM)

IZBIRA PODATKOV

Za namen razvrščanja ključnih besed bibliografskih enot raziskovalcev je treba na osnovi učnih podatkov (učne množice) izdelati ustrezni model. Uporabili smo različne vire, tako klasifikacije raziskovalnih dejavnosti, kot ključne besede iz baze SICRIS.

Pri izdelavi modela smo za osnovno učno množico uporabili klasifikacijo ARRS, in sicer seznam parov, opis klasifikacije in petmestno številčno kodo iz klasifikacije. Klasifikacija, predstavljena na straneh ARRS,[1] vsebuje okoli 500 parov šifra klasifikacije – opis. Prva številka petmestne šifre določa vedo, druga in tretja področje ter četrta in peta podpodročje.

Klasifikacija ARRS ima sedem ved:

- naravoslovje (1),
- tehnika (2),
- medicina (3),
- biotehnika (4),
- družboslovje (5),
- humanistika (6),
- interdisciplinarne raziskave (7).

Veda Interdisciplinarne raziskave za klasifikacijo ni primerna, ker gre za prepletanje ved, zato smo se odločili, da jo izločimo. S tem hkrati dosežemo enolično preslikavo v klasifikacijo CERIF,[2] ki je najbolj razširjena klasifikacija raziskovalne dejavnosti v Evropi.

Posamezna veda je razdeljena v povprečju na več kot deset področij. Področja so v povprečju razdeljena na dve podpodročji ali tri, vendar zelo neenakomerno. Določena področja sploh nimajo podpodročij. To je dodaten razlog za odločitev, da smo za klasifikator izbrali nivo področja.

Odločili smo se, da bomo posebej klasificirali angleška in slovenska besedila in rezultate med seboj primerjali.

Osnovno učno množico predstavljajo pari podatkov (besedilni opis – koda):

- opis področja in pripadajoča klasifikacija področja,
- opis podpodročja in pripadajoča klasifikacija področja.

Osnovna učna množica podatkov je majhna, 253 slovenskih in 253 angleških opisov področij in podpodročij s pripadajočo klasifikacijo. Za razširitev učne množice smo uporabili ključne besede raziskovalcev, raziskovalnih skupin, projektov in programov (v nadaljevanju ključne besede entitet) iz baze SICRIS in klasifikacije ARRS. Posamezne ključne besede smo upoštevali za vsako klasifikacijo posebej, če je entiteta imela dodeljenih več področij (v povprečju 1,5 klasifikacije ARRS na entiteto). Učna množica je tako vsebovala dodatnih 21.000 slovenskih in 19.000 angleških zapisov.

Za razvrščanje smo izbrali dve skupini besedil:

- klasifikacija raziskovalne dejavnosti CERIF – za 363 slovenskih in 363 angleških zapisov z namenom transformirati klasifikacijo ARRS v klasifikacijo CERIF;
- ključne besede bibliografskih enot raziskovalcev iz sistema COBISS, in sicer za 850.000 enot v slovenskem in 230.000 enot v angleškem jeziku z namenom klasificirati bibliografske enote.

PRIPRAVA PODATKOV

Podatke, namenjene tako učenju kot razvrščanju, smo pridobili iz baz SICRIS in COBISS ter jih združili. Uvozili smo jih s programom Oracle data miner v tabelo s petimi stolpci:

- Koda področja klasifikacije ARRS (stolpec CODE, podatkovni tip VARCHAR2 dolžine 5).
- Koda izvora podatkov ključnih besed (stolpec TYPE, podatkovni tip VARCHAR2 dolžine 3) za razdelitev podatkov glede na skupine področij in podpodročij v klasifikaciji ARRS; ključne besede raziskovalcev, projektov, programov, raziskovalnih skupin; področje klasifikacije CERIF, ključne besede v bibliografskih enotah.
- Koda jezika za slovenska oz. angleška besedila (stolpec LANG, podatkovni tip VARCHAR2 dolžine 3).
- Unikatna identifikacijska številka (ID); število, generirano kot funkcija kode jezika, kode izvora podatkov ter šifre entitete oz. kode klasifikacije (stolpec IDENT s primarnim ključem, podatkovni tip NUMBER dolžine 22).
- Besedilo (stolpec KEYWS, podatkovni tip VARCHAR2 dolžine 4000) – v primeru klasifikacije vsebuje polje opis klasifikacije, v drugih primerih ključne besede.

Posamezni stolpci tabele so v procesu podatkovnega rudarjenja poimenovani atributi. Cilj rudarjenja je preslikava ključnih besed iz bibliografskih enot v klasifikacijo ARRS za slovenski in angleški jezik posebej.

Atribut "Koda področja v klasifikaciji ARRS" smo izbrali kot ciljni atribut ter atribut "Besedilo" kot atribut, po katerem se izvede klasifikacija.

IZBIRA KLASIFIKACIJSKEGA ALGORITMA

V orodju Oracle data miner edino metoda podpornih vektorjev (Support Vector Machine) podpira razvrščanje besedil [1]. V primerjavi z drugimi algoritmi daje nadpovprečne rezultate, tako pri razvrščanju [5, 8] kot pri hitrosti izvajanja [6].

Klasifikacijski algoritem vsako besedilo predstavi kot vektor besed, v katerem so elementi obteženi glede na pogostost pojavitve besede v besedilu in glede na lastnosti medsebojne odvisnosti [5]. Postopek se izvaja v procesu učenja in pri procesu razvrščanja.

Zaradi hitrejšega izvajanja algoritma se v procesu učenja iz besedila izberejo najbolj reprezentativne besede, ki imajo pri razvrščanju največji pomen (angl. *feature selection*) [9]. Oblikujejo se podporni vektorji glede na posamezno kategorijo (klasifikacijo), tako da je med njimi maksimalna razlika.

V procesu razvrščanja besedila se za izdelan vektor besed preštejejo pojavitve besed, zbrane v podpornih vektorjih. Glede največje stopnje ujemanja s podpornim vektorjem se besedilo ustrezno razvrsti.

Pri nastavitvah parametrov modela smo večinoma ohranili privzete nastavitve [2]:

- 60 % podatkov je namenjenih izgradnji modela in 40 % testiranju.
- Vključeno je aktivno učenje, ki nadzoruje rast modela in optimizira čas izgradnje modela.
- Izbira odločitvene funkcije je avtomatska – sistem sam izbere med linearno in gaussovo funkcijo.
- Maksimalno število besed v besedilu je 50.
- Maksimalno število različnih besed za posamezno

kategorijo je 5.000.
- Za ciljno vrednost (angl. *target value*) pri izdelavi analiz smo izbrali področje Računalništvo in informatika.

IZDELAVA IN PRIMERJAVA MODELOV

Pripravili smo več različnih klasifikacijskih modelov in iskali najboljšega. Izbirali smo različne učne množice ob nespremenjenih nastavitvah parametrov modelov. Modele smo izdelali posebej za klasifikacijo angleških in slovenskih besedil.

Modele smo ocenjevali in primerjali glede na različne parametre. Spremljali smo:

1. Oceno modela, kot jo poda orodje Oracle data mining na osnovi testnih podatkov. Ocena pove, za koliko odstotkov je model boljši od t. i. modela Naive Bayes [2].
2. Oceno klasifikacije reprezentativnih besed po klasifikaciji ARRS, iz katere smo izbrali besede, tesno povezane s področjem. Seznam besed je predstavljal dodatno testno množico, kjer smo rezultate klasifikacij besed primerjali z želeno klasifikacijo. Oceno je predstavljal odstotek ujemanja rezultata klasifikacije z dejansko. Testno množico je sestavljalo 180 besed v slovenskem jeziku in 165 besed v angleškem jeziku.
3. Oceno transformacije vede v klasifikacijo CERIF. Klasifikacija CERIF in ARRS imata enake vede, le da ima klasifikacija CERIF medicino in biotehniko združeno v biomedicino. Ocena predstavlja odstotek ujemanja klasifikacij ved, kot jih je podal model glede na dejansko klasifikacijo. Testno množico je sestavljalo 363 opisov iz klasifikacije CERIF v slovenskem jeziku in 363 opisov v angleškem jeziku.
4. Oceno na osnovi ujemanja klasifikacije angleških in slovenskih besedil, ki pove, v koliko odstotkih je klasifikacija angleškega in slovenskega besedila enaka.
5. Oceno na osnovi odstotka ujemanja z vedo, pridobljeno iz števila UDK, ki je vezana na bibliografsko enoto.
6. Oceno razdelitve po vedah (v odstotkih) v primerjavi z razdelitvijo po številu UDK.

Modele smo zgradili na osnovi različnih učnih množic in vsakega posebej ocenili. Vse ocene modelov so predstavljene v tabelah.

Prva učna množica je vsebovala opise področij in podpodročij iz klasifikacije ARRS. Ocene modela so služile kot referenčna vrednost, na osnovi katere smo ocenjevali naslednje modele.

V drugo učno množico smo dodali ključne besede raziskovalcev, projektov, programov in raziskovalnih skupin (imenovane raziskovalne entitete). Učno množico smo tako povečali za 98 % ali dodatnih 40.000 zapisov. Kadar ima entiteta v bazi SICRIS dodeljenih več področij ARRS, se v učni množici iste ključne besede dodelijo za vsako klasifikacijo posebej. Model je po pričakovanjih dal slabše rezultate pri reprezentativnih besedah iz klasifikacije ARRS in boljše v primeru tansformacije vede in področja v klasifikacijo CERIF. Višja je bila tudi ocena ujemanja klasifikacije angleških in slovenskih besedil (tabela 1).

V tretjo učno množico smo dodatno vključili še opise iz klasifikacije CERIF. Na osnovi transformacije vede v klasifikacijo CERIF smo pridobili tudi transformacijo področij v klasifikacijo CERIF. Transformacijo smo pregledali in vnesli popravke. Na ta način smo pridobili dodatno množico za pridobitev ocene modelov. Opisom področij in podpodročij po klasifikaciji ARRS in CERIF smo želeli dati večjo težo. Ključne besede entitet se lahko ponavljajo (v povprečju vsaka 1,5-krat), po drugi strani pa dodeljena klasifikacija ARRS ni preverjena in je lahko tudi napačna. Da bi uravnotežili število skupin učnih podatkov (opisi področij in podpodročij iz klasifikacije ARRS, klasifikacije CERIF ter ključne besede entitet), smo opise področij in podpodročij ponovili v naboru učnih podatkov. Po nekaj različnih izbirah smo opise klasifikacije ARRS uporabili s šestkratno ponovitvijo ter opise klasifikacije CERIF s štirikratno ponovitvijo. Ocene so se pri vseh meritvah zelo izboljšale. Model smo ocenili kot primeren za klasificiranje ključnih besed iz bibliografskih enot in ga uporabili za kategorizacijo 850.000 ključnih besed iz bibliografskih enot, od tega jih je imelo 230.000 angleški prevod. Le v 44,80 % je prišlo do ujemanja klasifikacije slovenskih in angleških ključnih besed (tabela 2), ujemanje s klasifikacijo ved po UDK pa je bilo 61-odstotno v primeru slovenskega in 64-odstotno v primeru angleškega jezika (tabela 3).

V primeru četrte učne množice smo povsem zamenjali učne podatke. Iz klasifikacije ARSS smo poiskali besedne zveze oz. besede, ki enolično določajo klasifikacijo. Za oceno kvalitete izbranih besednih zvez smo izvedli primerjavo med klasifikacijo ključnih besed raziskovalcev, raziskovalnih skupin, projektov in programov, ter klasifikacijo, določeno glede na iskanje besednih zvez v ključnih besedah. Po večkratni primerjavi in izločitvi besednih zvez, ki klasifikacije ne določajo enoznačno, smo dosegli 73-odstotno ujemanje. Iz množice ključnih besed bibliografskih enot smo poiskali zapise, ki v začetnem delu (med prvimi 25. znaki) vsebujejo značilne besedne zveze. Z izločitvijo vseh duplikatov je učna množica vsebovala 88.000 slovenskih in 34.000 angleških zapisov. Izdelali smo

model in ga uporabili za kategorizacijo ključnih besed bibliografskih enot. Rezultati so bili v vseh ocenah slabši kot v predhodnem modelu, primerljivi so bili z rezultati drugega modela.

V primeru pete učne množice smo učni množici četrtega modela, pripravljeni na osnovi besednih zvez, dodali ključne besede entitet ter opise področij in podpodročij v klasifikaciji ARRS. V tem primeru je model dobil najboljše ocene. Pri oceni na osnovi ujemanja klasifikacije angleških in slovenskih besedil je bil rezultat v primerjavi s tretjim modelom izboljšan za okoli 5 % (tabela 2), podobno pri oceni na osnovi odstotka ujemanja z vedo, pridobljeno iz števila UDK (tabela 3).

Tabela 1: Ocene testnih primerov klasifikacij (ocene 1, 2, 3)

Model	Točnost modela	Ujemanje korenov	Ujemanje z vedo po CERIF
1 – slovensko	79,75 %	67,22 %	45,45 %
1 – angleško	82,98 %	74,55 %	61,16 %
2 – slovensko	47,05 %	63,33 %	65,84 %
2 – angleško	47,82 %	58,79 %	69,15 %
3 – slovensko	56,52 %	70 %	80,9 %
3 – angleško	58,87 %	80,60 %	80,9 %
4 – slovensko	94,41 %	62,78 %	59,78 %
4 – angleško	92,40 %	55,15 %	63,64 %
5 – slovensko	74,71 %	82,22 %	74,10 %
5 – angleško	66,48 %	80,61 %	75,21 %

Tabela 2: Ocena na osnovi ujemanja klasifikacije angleških in slovenskih besedil (ocena 4)

Model	Ujemanje vede (oba jezika)	Ujemanje področja (oba jezika)
3	65,80 %	44,80 %
4	56,47 %	33,52 %
5	70,83 %	50,71 %

Tabela 3: Ocene ujemanja vede po klasifikaciji bibliografskih enot z vedo, pridobljeno iz UDK-ja (ocena 5)

Model	Ujemanje z vedo UDK
3 – slovensko	60,78 %
3 – angleško	64,01 %
4 – slovensko	56,91 %
4 – angleško	60,33 %
5 – slovensko	67,47 %
5 – angleško	67,73 %

Tabela 4: Razdelitev bibliografskih enot po vedah glede na razdelitev po UDK

Model	Naravoslovje	Tehnika	Medicina	Biotehnika	Družbosl.	Humanistike
3 – slovensko	11 %	23 %	6 %	8 %	26 %	26 %
3 – angleško	14 %	20 %	8 %	6 %	22 %	30 %
4 – slovensko	29 %	18 %	4 %	5 %	25 %	18 %
4 – angleško	15 %	31 %	3 %	3 %	20 %	27 %
5 – slovensko	13 %	20 %	5 %	8 %	29 %	24 %
5 – angleško	15 %	23 %	4 %	6 %	24 %	29 %
Glede na UDK	15 %	18 %	11 %	6 %	29 %	21 %

UGOTOVITVE

Pri daljših besedilih, ki so vezana na različna raziskovalna področja, orodje za klasifikacijo ne omogoča, da bi dele besedila različno obtežili. Obtežitev smo z modelom 3 dosegli tako, da smo opise področij in podpodročij iz klasifikacij ARRS ter CERIF ponovili v naboru učnih podatkov in na ta način dosegli boljše rezultate v primerjavi z modelom 2 (tabela 1).

Najboljše rezultate smo dosegli v primeru, ko smo za učenje modela uporabili vse različne skupine učnih množic – model 5 v primerjavi z modelom 4 (tabele 1, 2, 3). Ker smo polovico podatkov pridobili z iskanjem korenov ključnih besed iz bibliografskih enot, smo na ta način tudi obtežili korenske izraze.

Glede na oceno odstotka ujemanja z vedo, pridobljeno iz števila UDK, je bila klasifikacija angleških besedil bolj točna (tabela 3), saj je orodje prilagojeno angleškemu jeziku, ki avtomatsko izloča veznike, predloge in druge besede, ki bi vnašale napako v model, pri slovenskih besedilih pa je klasifikacija težje izvedljiva zaradi sklonov samostalnikov.

Ocena razdelitve po vedah (v odstotkih) v primerjavi z razdelitvijo na osnovi števila UDK je dala najboljše rezultate za zadnji model, zgrajen na osnovi različih tipov učnih množic (tabela 4).

Za razdelitev bibliografskih del po raziskovalnih področij (v odstotkih) smo upoštevali povprečje angleških in slovenskih rezultatov zadnjega modela. Prvih deset mest je bilo razdeljenih med področja iz tabele 5.

Tabela 5: Prvih deset področij glede na število bibliografskih enot

Področje	Delež (%)
Ekonomija	5,38 %
Matematika	5,20 %
Jezikoslovje	3,79 %
Geografija	3,41 %
Vzgoja in izobraževanje	3,37 %
Etnologija	3,02 %
Literarne vede	2,95 %
Arheologija	2,87 %
Pravo	2,55 %
Rastlinska produkcija in predelava	2,47 %

Opombi

1 http://www.arrs.gov.si/sl/gradivo/sifranti/sif-vpp.asp

2 CERIF – the Common European Research Information Format je evropski šifrant raziskovalne dejavnosti.

Reference

[1] Oracle® Data Mining Concepts 11g Release 1 (11.1) (2008). Dosegljivo na: http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129.pdf.

[2] Robert Haberstro (2008). Oracle® Data Mining Tutorial. Dosegjivo na: http://www2.tech.purdue.edu/cit/Courses/CIT499d/ODMr%2011g%20Tutorial%20for%20OTN.pdf

[3] Oracle® Data Mining Administrator's Guide 11g Release 1 (2008). Dosegljivo na: http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28130.pdf.

[4] Oracle® Data Mining Application Developer's Guide11g Release 1 (2008). Dosegljivo na: http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28131.pdf.

[5] Milan Zorman, Vili Podgorelec, Mitja Lenič, Petra Povalej, Peter Kokol, Alojz Tapajner (2003). Inteligentni sistemi in profesionalni vsakdan. Maribor, Linea.

[6] Boriana L. Milenova, Joseph S. Yarmus, Marcos M. Campos (2005). SVM in Oracle Database 10g: Removing the Barriers to Widespread Adoption of Support Vector Machines. Dosegljivo na: http://www.vldb2005.org/program/paper/wed/p1152-milenova.pdf.

[7] Asa Ben-Hur, Jason Weston (2010). A User's Guide to Support Vector Machines. Dosegljivo na: http://pyml.sourceforge.net/doc/howto.pdf

[8] Fatimah Wulandini, Anto Satriyo Nugroho (2009). Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases. Dosegljivo na: http://asnugroho.net/papers/rict2009_textclassification.pdf.

[9] Tong Zhanng, Frank J Oles (2000). Text Categorization Based on Regularized Linear Classification Methods. Dosegljivo na: http://www.stat.rutgers.edu/home/tzhang/papers/ir01_textcat.pdf.

Spletne povezave

- SICRIS: http://sicris.izum.si/about/cris.aspx?lang=slv.
- ARRS: http://www.arrs.gov.si/sl/agencija/naloge.asp.
- COBISS: http://www.cobiss.net/platforma_cobiss.htm.
- Klasifikacija ARRS: http://www.arrs.gov.si/sl/gradivo/sifranti/sif-vpp.asp.
- Klasifikacija CERIF: http://www.arrs.gov.si/sl/gradivo/sifranti/sif-cerif-cercs.asp.
- Metoda podpornih vektorjev: http://en.wikipedia.org/wiki/Support_vector_machines.
- UDK: http://sl.wikipedia.org/wiki/Univerzalna_decimalna_klasifikacija.