# Učenje in odločanje: Analiza definicij osnovnih konceptov v Wikipedii z metodami analize besedil in omrežij

*Marko Bohanec*
Institut Jožef Stefan, Odsek za tehnologije znanja
Jamova cesta 39, SI-1000 Ljubljana, Slovenija
Tel: +386 1 4773309, e-mail: marko.bohanec@ijs.si
in Univerza v Novi Gorici, Vipavska 13, p.p. 301, SI-5000 Nova Gorica

**POVZETEK**

Prispevek predstavlja rezultate preliminarne analize besedil s področja učenja in odločanja. V ta namen smo analizirali 19 izbranih strani spletne enciklopedije Wikipedia z dvema računalniškima programoma: *Document Atlas* za analizo besedil (angl. *Text Mining*) in *Pajek* za analizo omrežij *(Network Analysis*). Pridobili smo informacije o vsebinski in strukturni povezanosti strani ter o odnosu med temeljnima proučevanima pojmoma: *učenje* in *odločanje*.

## 1 UVOD

Sodobna računalniška orodja za analizo podatkov (angl. *Data Mining*), analizo besedil (*Text Mining*) in analizo omrežij (*Network Analysis*) postajajo vse zmogljivejša, pa tudi vse bolj uporabna in dosegljiva. Še posebej so primerna za različne vrste iskanja in raziskovanja strukturnih in vsebinskih povezav med predmeti proučevanja. V tem prispevku predstavljamo primer uporabe dveh takšnih orodij za potrebe interdisciplinarnega raziskovalnega projekta. Pokazati želimo, da je mogoče na takšen način razmeroma hitro in preprosto dobiti zanimive in povsem uporabne raziskovalne rezultate. Ob tem želimo spodbuditi uporabo takšnih orodij tudi na drugih raziskovalnih področjih.

Projekt *Metodološki vidiki raziskovanja kognitivnih procesov: učenje in odločanje* (ARRS J7-9792, 2007–2009) je interdisciplinaren temeljni raziskovalni projekt, katerega cilj je razvoj interdisciplinarne raziskovalne paradigme, ki bi združevala in nadgrajevala dosedanje raziskovalne pristope (Kordeš, 2007). Projekt proučuje dva povezana kognitivna procesa, *učenje* in *odločanje*, ki ju poskuša bolje razumeti skozi integracijo interdisciplinarnih parcialnih znanj. V ta namen projekt združuje raziskovalce z različnih področij: filozofije, psihologije, nevroznanosti in umetne inteligence.

Motiv za izvedbo raziskave, opisane v tem prispevku, je nastal na enem od projektnih sestankov, kjer smo ugotovili, da je potrebno za učinkovito medsebojno delo in razumevanje natančno definirati slovar temeljnih pojmov s področja raziskovanja, jih opredeliti z različnih interdisciplinarnih vidikov in opisati povezave med njimi. Določili smo naslednje ključne pojme: *učenje*, *odločanje*, *misel*, *čustvo*, *dejanje*, *okolje*, *odgovornost*, *racionalnost*, *avtorstvo*, *znanje*, *cilji*. Sprva smo imeli namen slovar oblikovati sami, potem pa smo ugotovili, da so na spletni enciklopediji Wikipedia opisi teh pojmov (glejte dva primera na sliki 1) zelo dobri. Ustrezajo vsebinsko, strokovno in po obsegu, hkrati pa omogočajo različne analize povezav med njimi.

V okviru pričujoče raziskave smo si zato zastavili naslednje cilje:

- Ugotoviti, kateri ključni pojmi, pomembni za projekt, so opisani na Wikipedii in kako.
- Kako so ti pojmi med seboj vsebinsko povezani?
- Kakšen je odnos med ključnima konceptoma, "*učenje*" in "*odločanje*", ter katere so skupne točke med njima?

## 2 METODA DELA

Na spletni enciklopediji Wikipedia smo izbrali 19 strani (t.i. gesel oziroma člankov), ki so po našem mnenju najbolj ustrezale izbranim ključnim besedam: *Decision-making*; *Cognition*; *Choice*; *Learning*; *Memory*; *Behaviors*; *Experience*; *Education*; *Action_(philosophy)*; *Philosophy_of_action*; *Thought*; *Feeling*; *Environment*; *Context*; *Responsibility*; *Rationality*; *Knowledge*; *Objective_(goal)*; *Author*. Vse te strani smo zajeli dne 16.4.2008.

Besedila smo analizirali z dvema računalniškima programoma: *Document Atlas* in *Pajek*. Oba programa sta plod domačega znanja in brezplačna.

***Document Atlas*** (http://docatlas.ijs.si/) je program za klasifikacijo in vizualizacijo velikih količin dokumentov (Fortuna, et al., 2005). Program analizira množice besed, ki nastopajo v vhodnih dokumentih, ter razpozna tiste ključne besede, ki so med dokumenti najbolj podobne oziroma najbolj različne. Na tej osnovi oblikuje dvodimenzionalni prikaz dokumentov, kjer so med seboj podobni dokumenti prikazani skupaj. Grafično je prikazana je tudi gostota dokumentov. Dokumenti, ki so prikazani blizu skupaj, so med seboj vsebinsko povezani, bolj gosto posejani dokumenti pa običajno označujejo neko skupno temo oziroma obravnavani koncept.

V naši analizi smo *Document Atlas* uporabili na treh množicah dokumentov:

- originalni dokumenti, zajeti z Wikipedie v formatu HTML;
- prečiščeni dokumenti v tekstovnem formatu, pri čemer smo iz originalnih dokumentov izločili vse označevalne in nevsebinske dele (slika 2, levo);
- samo ključne besede, ki jih eksplicitno navaja vsak izvorni dokument (slika 2, desno).

***Pajek*** (http://pajek.imfm.si/) je program za analizo velikih omrežij (Batagelj, Mrvar, 2003; De Nooy et al., 2005). Omrežja so predstavljena v obliki grafov, ki jih mogoče analizirati, preblikovati in prikazovati s številnimi metodami, realiziranimi v programu.

V našem primeru smo analizirali graf povezav med spletnimi stranmi, upoštevajoč eksplicitne URL naslove, navedene na straneh. Za razliko od programa *Document Atlas*, s katerim smo iskali *vsebinske* povezave med dokumenti, gre pri *Pajku* torej za *strukturno* analizo spletnih povezav med dokumenti. Poleg osnovnih 19 strani smo v graf povezav vključili še vse tiste strani Wikipedie, na katere je mogoče priti s teh strani v enem koraku (skupaj 1777 dokumentov). Izdelali smo več grafičnih prikazov. V skladu z v uvodu omenjenimi cilji sta najpomembnejša dva prikaza:

- povezave med osnovnimi 19 dokumenti (slika 3), in
- graf povezav med dokumentoma *Decision-making* in *Learning* (slika 4).


**Slika 1**: Izseka dveh strani na spletni enciklopediji Wikipedia (Vir: Wikipedia, 2008)

## 3 REZULTATI

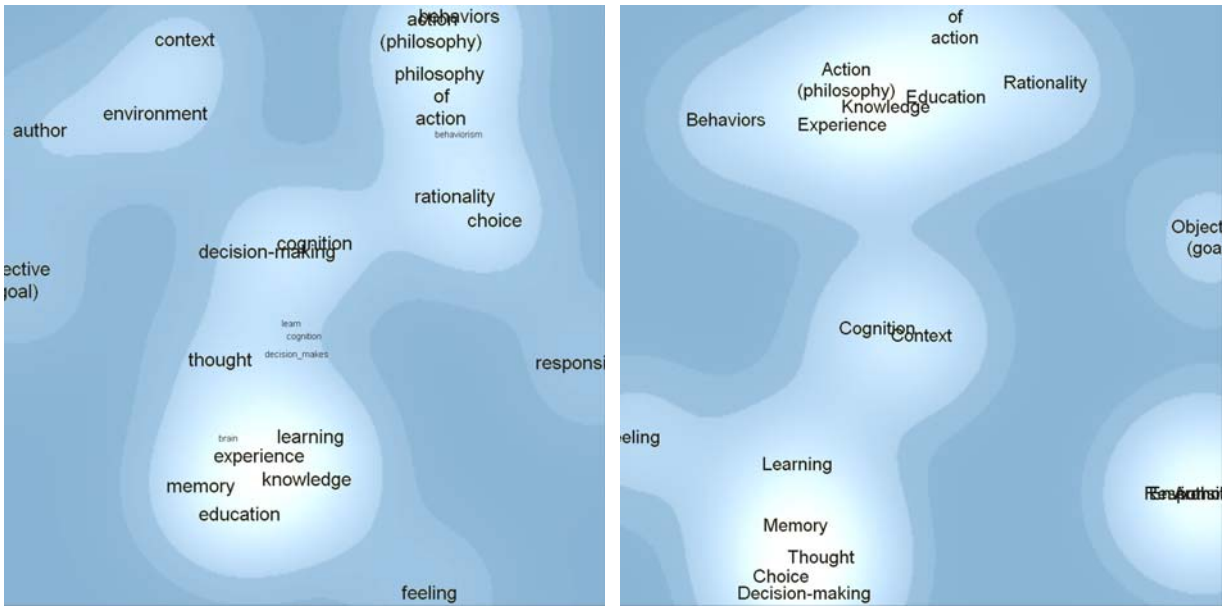
### 3.1 Analiza vsebinskih povezav s programom *Document Atlas*

Prva analiza, izvedena na originalnih dokumentih v formatu HTML, ni dala smiselnih rezultatov. Izkazalo se je, da nekateri elementi teh dokumentov, ki služijo le oblikovanju in označevanju besedila, motijo program, saj jih razume kot ključne besede. Prevladale so povezave med dokumenti, ko so bile bolj oblikovne kot vsebinske narave. Skupaj so se znašli dokumenti istih avtorjev, ki so uporabljali značilne oblikovne in slogovne elemente.

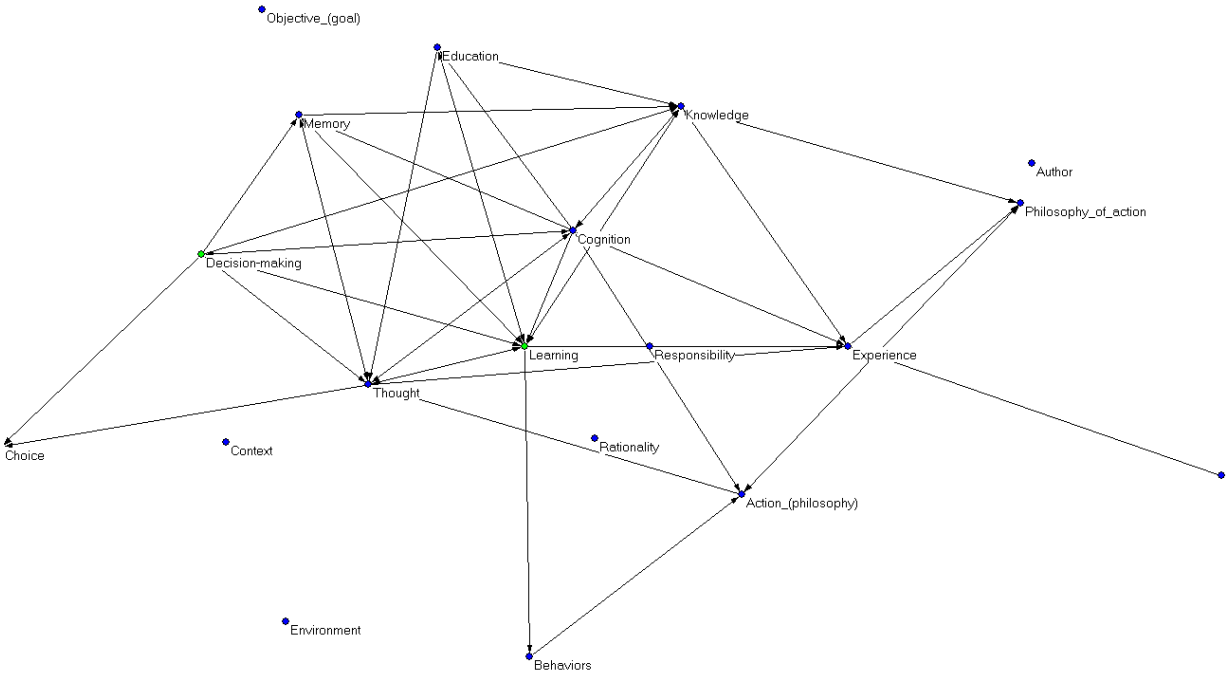
Boljši pa so bili rezultati na preostalih dveh množicah vhodnih dokumentov: prečiščenih besedilih (slika 2, levo) in seznamih ključnih besed (slika 2, desno). V obeh primerih sta se oblikovali dve večji skupini med seboj povezanih konceptov ter nekaj manjših skupin. Prvo večjo skupino označuje pojem *learning*, ki se povezuje s pojmoma *memory* in *thought* ter, v nekoliko večji razdalji, *decision-making*. Druga večja skupina vsebuje pojme s področja filozofije: *philosophy of action*, *behaviors*, *rationality*. Nekateri pojmi so enkrat bližje eni in drugič drugi skupini: *choice*, *education*, *experience*. Nekateri pojmi pa so vedno razmeroma oddaljeni in tvorijo svoje manjše skupine: *objective* (*goal*), *feeling*, *author*, *environment* in *context*.

### 3.2 Analiza strukturnih povezav s programom *Pajek*

Celotni analizirani graf je obsegal 1777 dokumentov, kar pomeni, da je 19 izbranih strani Wikipedie neposredno povezanih s 1758 drugimi dokumenti Wikipedie. Gre torej za veliko število povezav in zelo dobro povezanost izbranih dokumentov z drugimi dokumenti Wikipedie.



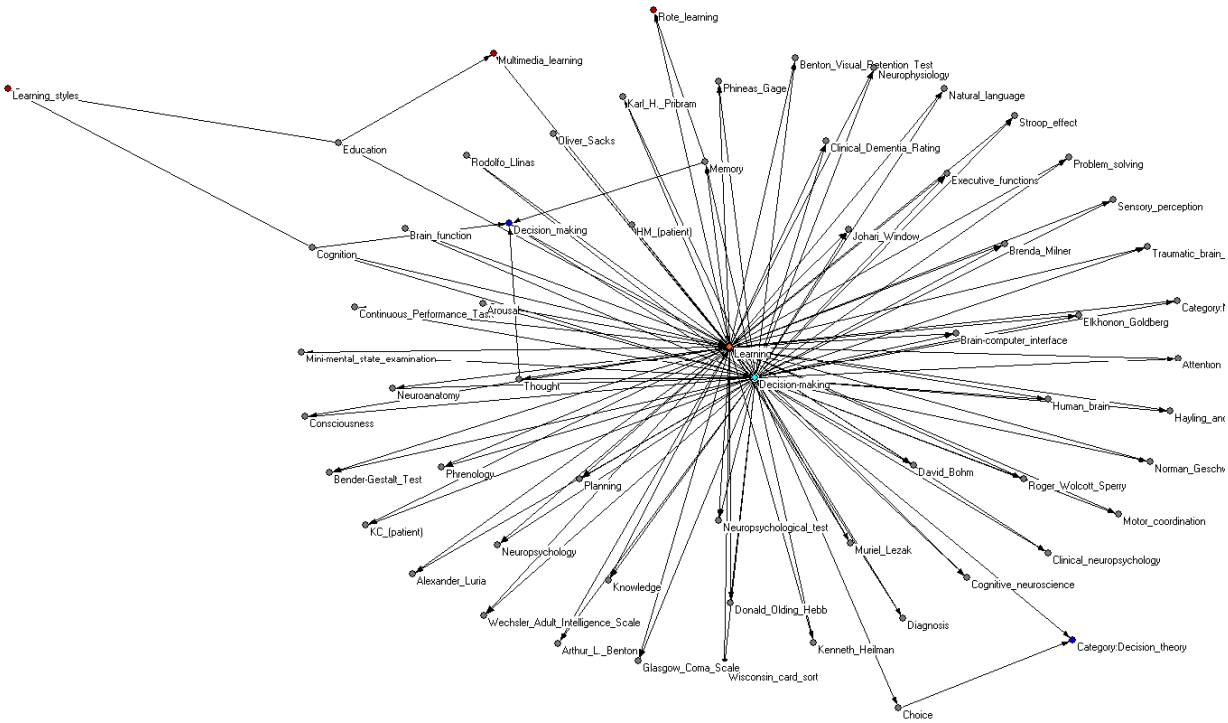

**Slika 2**: Vsebinska podobnost dokumentov: prečiščenih besedil (levo) in ključnih besed (desno)




**Slika 3**: Neposredne povezave med dokumenti

Neposredne povezave med izbranimi 19 dokumenti prikazuje slika 3. Puščice označujejo smeri povezav. Razvidna je osrednja vloga pojmov *learning*, *cognition*, *knowledge*, *thought*, *decision-making*, *memory*, *education* in *experience*, ki imajo veliko število povezav in so tudi med seboj dobro povezani. "Filozofska linija" *behaviors*, *philosophy_of_action* in a*ction_(philosophy)* je nekoliko odmaknjena, vendar tudi povezana med seboj. Tudi tu imamo elemente, ki so slabo povezani z ostalimi (*choice*, *feeling*) ali pa sploh ne (*context*, *environment*, *rationality*, *author*, *reponsibility*).

Zanimiv je tudi graf povezav med dokumentoma *learning* in *decision-making*. Slika 4 prikazuje ta dva dokumenta ter vse dokumente Wikipedie, preko katerih sta ta dva dokumenta povezana v največ dveh korakih. Jasno so razvidne ključne besede, ki povezujejo ta dva koncepta, na primer tiste, povezane s področji kognitivnih in nevroznanosti ter teorije in prakse odločanja. Veliko je povezav preko strani z opisi pomembnih raziskovalcev z omenjenih področij, kar je neposredna posledica dejstva, da analiziramo strani spletne enciklopedije.




**Slika 4**: Povezave med learning in decision-making v največ dveh korakih

## 4 ZAKLJUČKI

Opisana preliminarna raziskava je pokazala predvsem dvoje: da so sodobni programi za analizo besedil in omrežij resnično zelo zmogljivi, dostopni in preprosti za uporabo, ter da z njimi dobljeni rezultati omogočajo dober vpogled v vsebino in strukturo dokumentov z obravnavanega raziskovalnega področja. V našem primeru smo dobili zelo dober pregled nad področji, povezanimi z osrednjima proučevanima kognitivnnima procesoma: *učenje* in *odločanje*. Generirani grafični prikazi so sprožili več plodnih razprav v projektni skupini, pokazali so na povezave, ki bi jih sicer lahko spregledali, ter vzpodbudili razmišljanja o manjkajočih povezavah, ki bi jih morebiti lahko osvetlili v okviru projekta.

Pri delu s programom *Document Atlas* smo spoznali vpliv "prečiščenosti" besedil na rezultate analiz. Pred vsebinsko analizo je potrebno iz dokumentov izločiti označevalne in oblikovne elemente, sicer le-ti preveč vplivajo na rezultate. Pri vsebinski analizi smo dobili zelo dobre rezultate že samo iz ključnih besed, kar bi lahko poenostavilo morebitne nadaljnje analize na večjih množicah dokumentov.

Pri rezultatih programa *Pajek* nas je presenetilo veliko število povezav med stranmi Wikipedie. Izbranih 19 strani je neposredno povezanih s kar 1758 drugimi stranmi, kar pomeni, da število zunanjih povezav zelo hitro narašča z izborom osnovnih dokumentrov in da so le-ti dobro integrirani v spletno enciklopedijo. Po vsebinski plati smo ugotovili, da sta področji učenja in odločanja med seboj dobro povezani preko osnovnih strani, še predvsem preko strani *cognition* in *thought*. Ko v izbor vključimo še dodatne strani Wikipedie, še posebej izstopajo povezave med njima preko področij, povezanih z nevroznanostmi.

Opisana raziskava je bila preliminarna in v več pogledih omejena. Predvsem je šlo za subjektiven, *ad-hoc* izbor osnovnih dokumentov, ki bi bil prav lahko tudi drugačen. Število dokumentov, 19, je majhno in vsi izvirajo iz istega vira (Wikipedia). To pomeni, da je izbor pristranski, hkrati pa občutljiv na vplive avtorstva in oblikovanja dokumentov. V bodočih raziskavah bi bilo zato treba uporabiti bistveno večje število dokumentov iz različnih virov. V tem primeru bi bile smiselne tudi kvantitativne analize grafa povezav med dokumenti.

### Literatura


Batagelj V., Mrvar A. (2003): Pajek - Analysis and Visualization of Large Networks. in Jünger, M., Mutzel, P. (eds.) *Graph Drawing Software*. Berlin: Springer, 77–103.

De Nooy, W., Mrvar, A., Batagelj, V. (2005): *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

Fortuna, B., Grobelnik, M., Mladenić, D. (2005): Visualization of Text Document Corpus. *Informatica* 29, 497–502.

Kordeš, U. (2007): Projekt Metodološki vidiki raziskovanja kognitivnih procesov: učenje in odločanje kot poskus interdisciplinarnega raziskovanja. V: *Kognitivna znanost v Ljubljani* (ur. U. Kordeš, O. Markič), Evropski socialni sklad, Ljubljana, 129–136.

Wikipedia (2008) Wikipedia, the free encyclopedia. http://www.wikipedia.org/.