MOČ SODOBNEGA SPLETA

Marko Bajec
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Laboratorij za podatkovne tehnologije
Tržaška 25, 100 Ljubljana
marko.bajec@fri.uni-lj.si

SODOBEN SPLET

Količina podatkov, ki jih vsakodnevno ustvarimo in objavimo na spletu, je ogromna. Zadnje statistike kažejo, da na družabno omrežje *FaceBook* vsak dan odložimo 500 tera bytov vsebin, na *Youtube* 72 ur video posnetkov, obseg podatkov celotnega spleta pa se menda dnevno poveča kar za milijon tera bytov ali en exa byte (Baeza-Yates, 2011). V tako ogromni količini podatkov najdemo vse mogoče – skoraj da ni več teme, ki ne bi bila na spletu tako ali drugače prisotna. Pomemben prispevek k temu je bila digitalizacija medijev, saj se danes na spletu objavi praktično celotna dnevna kronika, vključno z osebnimi vidiki in interpretacijami piscev ter komentarji bralcev.

Povsem nove priložnosti se pojavljajo z revolucijo on-line socialnih medijev. Številni sociologi jim pripisujejo neizmerno moč, saj omogočajo dostop do virtualnih skupnosti, ki so po številu članov večje od populacije največjih držav sveta. Internet tako ni več le infrastruktura, prek katere dostopamo do shranjenih podatkov, temveč gre za medij, prek katerega imamo dostop do ogromnega števila ljudi, njihovih mnenj, razmišljanj, osebnih pogledov. Kot primer, znane osebnosti, kot so Justin Bieber in Lady Gaga, lahko prek Twitterja naslovijo več svojih "sledilcev" kot pa na primer Angela Merkel v katerem izmed svojih javnih nastopov. Še več, prek socialnih medijev na mnenja posameznikov lahko tudi namensko vplivamo (Bik in Goldstein, 2013).

Današnja podjetja (ter druge organizacije) se omenjenega dobro zavedajo in poskušajo uporabiti splet kot komplementaren podatkovni vir, iz katerega poskušajo izluščiti tisto, kar jim je v pomoč pri njihovem delovanju. V nadaljevanju bomo predstavili nekaj zanimivih uporab spleta, od tega, kako razna podjetja sledijo dogajanju na trgu do tega, kako s pomočjo spletno dostopnih podatkov spoznavajo svoje stranke in posledično večajo svojo prodajo. Predstavljen bo zanimiv fenomen, iz katerega izhaja, da so množice v določenih okoliščinah »pametnejše« od elitnih posameznikov (ang. »*The Wisdom of The Crowds*«) ter da je splet lahko dober simulator množic.

V zadnjem delu prispevka bodo predstavljeni izzivi, s katerimi se soočamo pri avtomatski obdelavi podatkov s spleta. Iskanje prek spletnih brskalnikov in prebiranje zadetkov je namreč izjemno omejeno, saj lahko na tak način v nekem razumnem času obdelamo le izjemo majhen del relevantnih vsebin. Avtomatski zajem in interpretacija spletnih podatkov pa je po drugi strani zahtevna, saj so podatki na spletu objavljani v zelo različnih in nestrukturiranih oblikah. Na kratko bomo omenili dve tehnologiji, ki se v ta namen uporabljata: ekstrakcija podatkov s spleta (Web Data Extraction) ter spletno rudarjenje (Web Data Mining).

PRIMERI UPORABE

Agilno določanje cen

Lastniki svetovno znane veriga motelov z nekaj tisoč lokacij po vsem svetu so ugotovili, da se število on-line rezervacij prek spletnih strani podjetja zmanjšuje, čedalje več provizije pa plačujejo posrednikom – agencijam – saj uporabniki v lovu za najboljšimi cenami skoraj vedno najdejo agencijo, ki ima v tistem trenutku posebno akcijo. S tem se je zmanjševal obisk spletnih strani podjetja ter skupno število zakupov, saj so stranke na spletnih straneh agencije iskale med vsemi možnostmi in ne le med moteli tega podjetja.

Z vpeljavo posebne rešitve, ki je podjetju omogočala avtomatski zajem podatkov o ponudbah za namestitve, ki jih v nekem trenutku strankam ponujajo agencije (cene, storitve, paketi, posebne ugodnosti), so lahko v podjetju oblikovali model določanja ponudbe, s katerim so se približali kupcu. V enem letu so povečali prisotnost podjetja na trgu ter prodajo za 15%.

Sledenje spletni prodaji

Podoben problem je imel znan evropski proizvajalec fotografskih aparatov in pripomočkov. Za pametno določanje cen so potrebovali dober pregled konkurenčne ponudbe. Dodatno skrb so jim povzročali preprodajalci, ki so pogosto prodajali stare komponente za nove ter imitatorji, pri katerih so se pod njihovo blagovno znamko znašli neoriginalni fotografski aparati in oprema.

V podjetju so razvili posebno rešitev za zajem podatkov s spleta, ki jim omogoča slednje praktično vsaki prodaji njihovih produktov. Z uporabo šifranta serijskih številk originalnih delov lahko brez težav identificirajo kršenje poslovne politike ali nelegalno prodajo in sprožijo ustrezna ukrepe.

Bogatenje podatkov o strankah

Bogatenje podatkov o strankah je posebno popularen način uporabe spleta kot komplementarnega vira podatkov. Na spletu je namreč moč najti veliko dodatnih podatkov, ki nam o strankah, s katerimi imamo opravka, veliko povedo. Podjetja se danes tega pogosto poslužujejo in svoje podatkovne zbirke dopolnjujejo s podatki, ki jih o posameznikih najdejo na spletu. To jim omogoča, da stranke bolje spoznajo in jim posledično približajo svojo ponudbo. Pojavljajo pa se tudi podjetja, ki bogatenje podatkov ponujajo kot svojo storitev. Največ te ponudbe je na področju Združenih držav Amerike. Podatki, ki jih ponujajo, vključujejo različne kategorije: osebni, geografski, demografski, psihografski, socio-ekonomski itd.

Podatkovna fuzija

Zanimiv primer uporabe spleta kot bogatega vira podatkov smo demonstrirali tudi v Laboratoriju za podatkovne tehnologije Fakultete za računalništvo in informatiko. Razvili smo t.i. *Supervizor+* (po vzoru aplikacije, ki je nastala v okrilju Protikorupcijske komisije RS).

Supervizor+ je spletna računalniška rešitev za iskanje povezav med poslovnimi subjekti in fizičnimi osebami v Sloveniji. Povezave so vizualizirane s pomočjo grafa, kjer vozliča predstavljajo pravne ali fizične osebe, povezave med vozlišči pa predstavljajo eno izmed naslednjih pomenskih odvisnosti: poslovne odvisnosti (lastnik, solastnik, družbenik,

delojemalec, plačnik, prejemnik…), politične odvisnosti (pripadnik/član, simpatizer, nasprotnik), družbene odvisnosti (sorodnik, prijatelj, registriran na istem naslovu…).

S pomočjo povezav, ki so vzpostavljene med subjekti, lahko učinkovito pregledujemo različne vrste skupnosti (večje število tesno povezanih subjektov) ter druge neobičajne vzorce (**Napaka! Vira sklicevanja ni bilo mogoče najti.**, **Napaka! Vira sklicevanja ni bilo mogoče najti.**). Sistem omogoča tudi iskanje povezav med poljubnima dvema subjektoma. Če takšna povezava ali več povezav obstaja, sistem vizualizira del grafa, ki vsebuje omenjene povezave. Primer ekrana Supervizorja+ prikazuje slika spodaj.

Gre za primer fuzije podatkov iz različnih podatkovnih virov, ki so spletno dostopni:

- Strukturirani poslovni viri (AJPES, UJP, ZZZS, Supervizor, …)
- Strukturirani spletni viri (Wikipedia ipd.)
- Socialna in profesionalna omrežja (Facebook, Google+, LinkedIn itd.)
- Spletni časopisi (Finance, Dnevnik, Delo itd.)
- Drugo (prvih nekaj zadetkov, pridobljenih z uporabo meta-iskalnika).

Modrost množic

Modrost množice je proces oblikovanja mnenja, kjer namesto ekspertnega znanja upoštevamo mnenje širše množice ljudi. Temelji na predpostavki, da ko gre za vprašanja, povezana s splošnim znanjem, kvantitativnim ocenjevanjem ali prostorskim sklepanjem, so odgovori množic vsaj tako dobri, če ne boljši kot odgovori posameznih ekspertov. Pogosta intuitivna razlaga za ta fenomen je, da je pri odgovorih posameznikov prisoten šum, ki se z agregacijo večja števila odgovorov in povprečenjem preprosto izniči.

V poslovnem smislu je ta proces prvi opisal James Surowiecki v svoji knjigi »*The Wisdom of the Crowds*« (Castilo, 2013). V njej navaja zanimiv primer iz leta 1906, ko je statistik Francis Galton na kmečkem sejmu v Plymouthu (Velika Britanija) dal osemsto ljudem oceniti, koliko tehta bik, ki se je na sejmu prodajal. Zmagal je mesar, ki je imel očitno dober občutek, za Galtona pa je bilo pomembnejša ugotovitev, da je povprečje, ki ga je izračunal iz osemsto odgovorov, predstavljalo še bistveno boljšo oceno – od resnične vrednosti se je razlikovalo le za slab procent. Kasneje je bilo v okviru kognitivne znanosti pokazano, da lahko mnenje množic modeliramo z verjetnostno distribucijo, pri kateri je povprečna vrednost zelo blizu resnične vrednosti, ki se meri. V svoji knjigi Surowiecki na številnih primerih pokaže, da ta fenomen velja na več področjih, primarno v ekonomiji in psihologiji.

Danes ta fenomen izkoriščajo mnoga podjetja, med drugimi Google, Wikipedia, Flickr, Yahoo! Answers ipd. S tem povezano je tudi področje, poznano kot skupinska inteligenca (ang. *Collective Intelligence*). Znanje, razumevanje oziroma v širšem pomenu inteligenca se ne oblikuje le v možganih posameznika, temveč tudi v skupinah posameznikov, ki skupaj delujejo, preprosto rečeno, na pameten način. Splet, kot ga poznamo danes, daje temu fenomenu izjemno moč. Ključno vprašanje seveda je, kako povezati ljudi in računalnike, tako, da bodo skupinsko delovali boljše, pametneje kot katerikoli posameznik, skupina ali računalnik kadarkoli prej (Bothos, Apostolou, Mentzas, 2012).

TEHNOLOGIJE

Uporaba spleta, kot omenjeno v prejšnjih razdelkih, zahteva avtomatizirano obdelavo, kar vključuje identifikacijo, filtriranje, zajem ter interpretacijo posameznih podatkov. Področje, ki se s tem ukvarja, je ekstrakcija podatkov s spleta (ang. Web Data Extraction ali Web scraping). Gre za postopek, kako iz nestrukturiranih spletnih virov (tipično HTML) pridobiti podatke, ki na zanimajo, ter jih pretvoriti v strukturirano obliko, ki bo omogočala nadaljnjo računalniško obdelavo (Glez-Peña e tal, 2013).

Za potrebe ekstrakcije podatkov s spleta se uporabljajo zelo različne tehnike, od uporabe regularnih izrazov, XML analize (parsanja) DOM dreves, povpraševalnih jezikov (XPath, XQuery), podatkovnega rudarjenja ipd.

Spletno rudarjenje se uporablja tudi širše, ne le za namen ekstrakcije podatkov temveč tudi za potrebe identifikacije vzorcev na spletu. V to skupino sodi rudarjenje po podatkih uporabe spleta (ang. Web Usage Mining), rudarjenje vsebine (ang. Web Content Mining) ter rudarjenje strukture (ang. Web Structure Mining) (Singh in Kaur, 2013).

LITERATURA

Bothos E., Apostolou D., Mentzas G. (2012), Collective intelligence with web-based information aggregation markets: The role of market facilitation in idea management Original Research Article, Expert Systems with Applications, Volume 39, Issue 1, January 2012, Pages 1333-1345

Baeza-Yates R. (2011), An Introduction to Web Retrieval, ESSIR 2011, Koblenz, Germany

Bik H. M., Goldstein M. C. (2013), An Introduction to Social Media for Scientists. PLoS Biol 11(4): e1001535. doi:10.1371/journal.pbio.1001535

Castillo, M. (2013), The Wisdom of Crowds, American Journal of Neuroradiology.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. Briefings in bioinformatics.

Singh, M., & Kaur, N. (2013). A Review on Various Web mining Techniques with Purposed Algorithm of K-means Web Ranking.