Google Flu Trends

Urh Peček

14 3 2021

Opis ideje Google Flu Trends

Epidemije sezonskih grip vsako leto povzročijo med 250.000 in 500.000 smrti zaradi bolezni dihal in tako predstavljajo veliko skrb za javno zdravje. Z zgodnjim odkrivanjem bolezni in hitrim odzivom pa lahko pomembno zmanjšamo vpliv sezonske in pandemske gripe. Tradicionalni nadzorni sistemi za nadzor in preprečevanje bolezni v ZDA (CDC) tedensko objavljajo podatke o obiskih zdravnika zaradi simptomov podobnih gripi vendar pa se podatki običajno objavljajo z 1 do 2 tedenskim zamikom. Ker pa zelo veliko američanov na spletu dnevno išče informacije glede bolezni je eden od načinov za izboljšanje zgodnjega odkrivanja prisotnosti bolezni lahko spremljanje iskanja zdravstvenih nasvetov v spletnem brskalniku milijonov uporabnikov po spletu. Raziskano je bilo, da so iskanja v zvezi z boleznijo močno povezana z obiski zdravnika zaradi simptomov podobnih gripi in tako omogočajo relativno natančno oceno trenutne ravni razširjetnosti gripe. Google Flu Trends, krajše GFT je bila spletna storitev upravljana s strani Googla katere prvotna motivacija je bila, da bi lahko z zgodnjim prepoznavanjem aktivnosti bolezni in hitrim odzivom zmanjšali vpliv sezonske in pandemične gripe. Ideja je bila ta, da bi s spremljanjem vedenja uporabnikov na spletu in združevanjem iskalnih poizvedb, Google poskušal napovedati dejavnost gripe v časovni in prostorski komponenti. Ker Google search omogoča enostavno obdelavo podatkov bi bile lahko ocene razširjenosti gripe dosledno 1 do 2 tedna prej dostopne kot nadzorna poročila o gripi. Sprva so se osredotočili na devet regij v ZDA kasneje pa so v več kot 25 državah tedensko izračunavali stanje prisotnosti gripe z zamikom poročanja približno en dan. Model seveda ni bil ustvarjen kot nadomestilo tradicionalnih nadzornih sistemov vendar zgolj kot njihovo dopolnilo, ki jim omogoča pravočasen in primeren odziv na sezonske epidemije.

Opis modela in načina zbiranja podatkov

Predlagani sistem odkrivanja razširjenosti gripe temelji na avtomatiziranih metodah iskanja besednih zvez povezanih z gripo vtipkanih v spletni brskalnik. Med letoma 2003 in 2008 je bilo obdelano več sto milijard posameznih iskanj in z njihovim združevanjem je bila izračunana časovna vrsta tedenskega štetja 50 milijonov najpogostejših iskanj v ZDA. Razvit je bil model, ki ocenjuje verjetnost, da je naključni obisk zdravnika povezan z gripo. Zapisan je bil kot $logit(I(t)) = \beta_0 + \beta_1 logit(Q(t)) + \epsilon$, kjer je $I(t)$ delež obiskov zdravnika povezanih z gripo in $Q(t)$ delež spletnih iskanj povezanih z gripo v času t ter $logit(p) = ln(\frac{p}{1-p})$. Za izgradnjo modela so bili poleg spletnih poizvedb uporabljeni javno dostopni podatki poročanja CDC o povprečnem deležu tedenskih ambulatnih obiskov, povezanih z gripo. Ugotovljeno je bilo, da upoštevanje n=45 iskalnih poizvedb zagotavlja najboljše napovedi modela in vseh teh avtomatično izbranih 45 poizvedb je bilo povezano z gripo.

Big data

Presodimo ali gre v primeru Google Flu Trends za množične podatke, tako imenovane “big data”, kar leti predvsem na njihovo ogromno količino. Za izgradnjo modela je bilo za določitev upoštevanih iskanih besednih zvez izmerjeno kako učinkovito bi se podatki CDC o gripi v določeni regiji ujemali z rezultati modela, če bi kot edino spremenljivko uporabili delež upoštevanih spletnih iskanj. V postopku izbire so bili uporabljeni vsi tedni med 28. septembrom 2003 in 11. marcem 2007. Dodatnih 42 tednov med 18. marcem 2007 in

11. majem 2008 pa je bilo uporabljeno za potrditev izbire iskanih besednih zvez. Vsako od obravnavanih 50 milijonov najpogostejših spletnih iskanj je bilo uporabljeno v namen testiranja katere besedne zveze bi najbolj natančno modelirale delež obiskov povezanih z gripo v posamezni regiji. Skupno je bilo uporabljeno 450 milijonov različnih modelov za testiranje kandidatov upoštevanih spletnih iskanj, za delo pa je bilo uporabljenih več sto računalnikov. Zahtevnost bi sicer lahko zmanjšali s filtriranjem spletnih iskanj povezanih zgolj z gripo, vendar bi lahko agresivno filtriranje povzročilo izgubo dragocenih podatkov, poleg tega pa bi morda izbrane poizvedbe, ki niso povezane z gripo nakazale, da je bil pristop njihove izbire nepravilen. Na zgoraj opisan mehanizem pridobivanja podatkov ter razvoj modela hitro vidimo, da imamo zadostno količino argumentov, da lahko ideji Google Flu Trends v opis dodamo besedno zvezo big data.

Dobre in slabe plati modela ter točnost podatkov

Hitro so se pojavila vprašanja o zasebnosti, katerim se je Google Flu Trends skušal izogniti tako da, nobenih spletnih iskanj ni bilo mogoče povezati z določenim posameznikom in zbirka podatkov ni vsebovala identitete uporabnika, ki je izvedel iskanje ali njegove fizične lokacije. Je pa bil v namen odkrivanja iz katere regije je iskanje prišlo zabeležen njegov internetni protokol.

V začetni fazi so bile napovedi Googla 97% natančne v primerjavi s podatki CDC, vendar pa so kasnejše napovedi Googla, zlasti med letoma 2011 in 2013 stalno precenjevale relativno incidenco gripe. Ena izmed težav je ta, da ljudje, ki po spletu iščejo podatke o bolezni ponavadi zelo malo vedo o diagnosticiranju gripe in raziskujejo simptome bolezni, ki so podobni gripi vendar pa dejansko to niso. Poleg tega so možnosti za najdbo poizvedb, ki ustrezajo gripi vendar so z njo nepovezane zelo velike. Googlov algoritem je bil precej ranljiv za prekomerno prilagajanje sezonskim izrazom, ki niso povezani z gripo, na primer "srednješolska košarka". Tudi panika ob domnevno novem sevu lahko povzroči prekomerno spletno iskanje povezano z boleznijo in tako pretirane ocene tekoče razširjenosti, saj upoštevanih spletnih iskanj ne uporabljajo samo posamezniki z upoštevanimi simptomi. Težava je bila tako v tem, da je model lahko meril samo tisto, kar ljudje iščejo, v postopku pa se ni analiziralo, zakaj so bile besede zares iskane. Z odstranitvijo človeškega vložka in dovoljevanjem obdelave neobdelanih podatkov je model napovedoval golj na podlagi iskalnih poizvedb iz prejšnjih let. Enega izmed večjih problemov predstavlja tudi nenehno posodabljanje Googlovega algoritma iskanja ter njegovih predvidevanj iskanj, saj lahko s funkcijo autosuggest ljudi dodatno vzpodbudi v iskanje zadetkov povezanih z gripo. Dodatno pa je Google lahko ob iskanju različnih bolezni bolj nagnjen h ponujanju rezultatov povezanih z gripo. Da bi se nenatančnim napovedim lahko izognili bi bilo potrebno redno posodabljanje uporabljenih modelov, brž ko bi zaznali njihovo nenatančnost. Vendar pa je to odkrivanje napak precej oteženo saj se napake ne odkrivajo zlahka temveč z izkušnjami.

Problematično je tudi združevanje big in small data. Ugotovljeno je bilo, da lahko z združevanjem podatkov GFT in zaostalih CDC izboljšamo le enega izmed GFT ali CDC. Google niti nikoli ni razkril katerih 45 iskalnih izrazov je uvrstil v algoritem, niti kako jih je utežil za ustvarjanje napovedi, kar je povezano s tem, da Google ne pojasni svojih algoritmov in tako blokira znanstvene raziskave, ki bi le te lahko izboljšale.

Ključna spoznanja in "aha" momenti

- Vrednost podatkov, ki jih imajo subjekti, kot je Google, je skoraj neomejena, če se pravilno uporabljajo. To pomeni, da so korporativni velikani, ki imajo te podatke, odgovorni za njihovo uporabo v najboljšem interesu javnosti. Korporacije in potrošniki so del širše družbe in mnogi od arhivov velikih podatkov ponujajo vpoglede, ki bi lahko koristili prav vsem.
- Čeprav so bila Googlova prizadevanja za napovedovanje razširjenosti gripe dobronamerna, so bila v pogledu metode in podatkov izjemno nepregledna, zaradi česar je bilo nevarno, da se pri sprejemanju odločitev zanese na Google Flu Trends. Zato bi bilo ključno, da bi Google v prihodnosti, v kolikor bi se ponovno lotil podobnega projekta, vseskozi posodabljal modele, saj vrednost podatkovnega toka kot smo videli lahko hitro upade.
- Zakonski del hranjenja podatkov o uporabnikovih spletnih poizvedbah je običajno precej spregledan, saj večina ljudi seveda ne prebere prebere določil in pogojev, vendar pa je to nekakšna kupčija obeh

strani, saj je posameznik deležen storitev, družba pa nekaj podatkov.

- Ključno pri big data je tudi to, da imajo imetniki velikih podatkov občutljivih uporabnikov možnost njihove delitve v znanstvene namene ter obenem ohranijo identiteto uporabnikov skrito ter podatke popolnoma varne. Tako se pojavi ključno vprašanje, kako nadgraditi in okrepiti prizadevanja ščitenja občutiljivih podatkov ter varovati zasebnost posameznikov ter obenem zadostiti lastniškim interesom imetnikov velikih podatkov.