Filternet: Sistem za tematsko filtriranje, pregledanje i arhivu podataka

Dušan Josipović[1], Dušan Vujošević[2]

***Sadržaj* — Upravljanje ličnim znanjem je skup kognitivnih i tehnoloških aktivnosti koje korisnik svakodnevno obavlja da bi efikasno prikupio, klasifikovao, sistematizovao, sačuvao, pretražio, pregledao i podelio informacije. Uz sve veći broj kreiranih digitalnih sadržaja taj proces postaje i sve kompleksniji. Kao odgovor na kompleksnost, u ovom radu se predstavlja sistem za tematsko arhiviranje i pretraživanje podataka, sa fokusom na veb stranice, a koji olakšava upravljanje ličnim znanjem. Svrha predstavljenog sistema je pomoć pri istraživanju radi sticanja znanja u kome je potrebno usvojiti veći broj međusobno povezanih pojmova. Glavni procesi predstavljenog sistema su skrejpovanje veb stranica i embedovanje kojim se dobija interaktivna mapa za prikaz tematskih oblasti. U radu se konceptualno prikazuju komponente sistema i analizira se primer konkretne implementacije. Ostvareni rezultat je praktična implementacija za upotrebu i dalje izučavanje.**

***Ključne reči* — Filtriranje, Filternet, Pronalaženje informacija, Pretraživači, Nenadgledano obučavanje**

I. UVOD

Razvoj računarstva i interneta omogućio je protok do sada neviđene količine informacija između ljudi sa svih strana sveta. Iako pretraživači (npr. Google, Bing) nastali pre nekoliko decenija znatno olakšavaju pronalaženje informacija na internetu, ovaj zadatak, u zavisnosti od traženih informacija, može biti veoma izazovan. Budući da se količina podataka povećava eksponencijalno, može biti sve teže pronaći relevantne rezultate pretrage. Kao nova tehnička

[1] Dušan Josipović, Beograd, Srbija (email: josipovic@fastmail.com)

[2] Dušan Vujošević, Računarski fakultet, Beograd, Srbija (email: dvujosevic@raf.rs)

pomoć upravljanju znanjem, u ovom radu se predstavlja sistem za efikasno pregledanje, pronalaženje i arhiviranje dokumenata. Primarni cilj upotrebe predstavljenog sistema je pomoć pri učenju i istraživanju određene oblasti, onda kada je potrebno usvojiti netrivijalan broj međusobno povezanih koncepata iz određenog polja i skladištiti ih radi ponovnog pristupa.

Čovek i računar

Simbioza čovek-računar je već uveliko omasovljena. Računari (pametni telefoni, laptopovi itd.) skladište sve veće količine podataka koje ljudski mozak ne može pouzdano sačuvati. Trenutni stepen razvitka tehnologije ne dozvoljava dublju integraciju koja bi omogućila veći protok informacija ljudskog mozga sa jedne strane i mašine sa druge, pa i percepcija da živimo u simbiozi sa računarima nije učestala. Brzina upisivanja podataka i upravljanja mašinom je ograničena brzinom kucanja na tastaturi, a čitanje ograničavaju brzina ljudskog čitanja ili slušanja, iako je ljudski mozak daleko brži u generisanju misli. Teorijski gledano, sistem koji se u ovom radu predlaže je zamišljen da efikasno poboljša procese usvajanja i organizacije koncepata, te njihovo ponovno pronalaženje uz pomoć računarskog dela simbioze.

Ciljevi rada

Prva pomisao pri pronalasku objašnjenja nekog pojma je pretraga na internetu. U zavisnosti od stepena poznavanja oblasti kojoj pojam pripada, konteksta i pogleda na svet, različita objašnjenja će odgovarati različitim korisnicima. Stoga će pristup „isti rezultati za sve korisnike“ biti tek osrednjeg kvaliteta. Jedan od ciljeva ovog rada je da se omogući korisnicima da imaju preferencije u izboru izvora informacija. Pre nego što pretraže željeni pojam sa generalnom pretragom (npr. Google.com) moćiće da pristupe svemu što su njihovi probrani izvori napisali.

Komercijalizacija interneta dodatno otežava zadatak pretrage. Finansijska nagrada za oglašivače koji uspeju da skrenu pažnju na svoje veb stranice dovoljna im je motivacija da dovode u zabludu pretraživače o sadržaju koji se nalazi na reklamiranom veb sajtu. Sa druge strane, i pretraživači koji zarađuju od reklama nemaju u potpunosti usaglašen cilj sa korisnicima njihove pretrage. Dok korisnici tragaju za kvalitetnim informacijama, pretraživači zarađuju samo od klikova na reklame. Pretraživači takođe moraju da paze na kvalitet rezultata, ali prave kompromis sa zaradom koju ostvaruju.

Čak i nakon 20 godina postojanja kao najviše korišćeni pretraživač, iako je bilo znatnog napretka, Google.com i dalje ima problema da efikasno kontroliše manipulaciju optimizacije sajtova za pretraživače. Iako se tehnike menjaju, danas je moguće primetiti klonove poznatih veb portala prikazanih u pretrazi, sa kompletno kopiranim sadržajem, iznad originalih izvora tih podataka.

Osim pronalaženja tekstualnih informacija na internetu njihovo efikasno skladištenje radi kasnijeg pronalaženja je nezgrapno (npr: eng. Bookmarks) ili zahteva prilagođen softverski sistem za tu namenu (npr: lokalnu instancu Viki softvera) u koji bi korisnik pohranjivao podatke. Količina informacija kojoj se korisnici izlažu na dnevnom nivou je sve više opterećujuća za efikasno upravljanje znanjem.

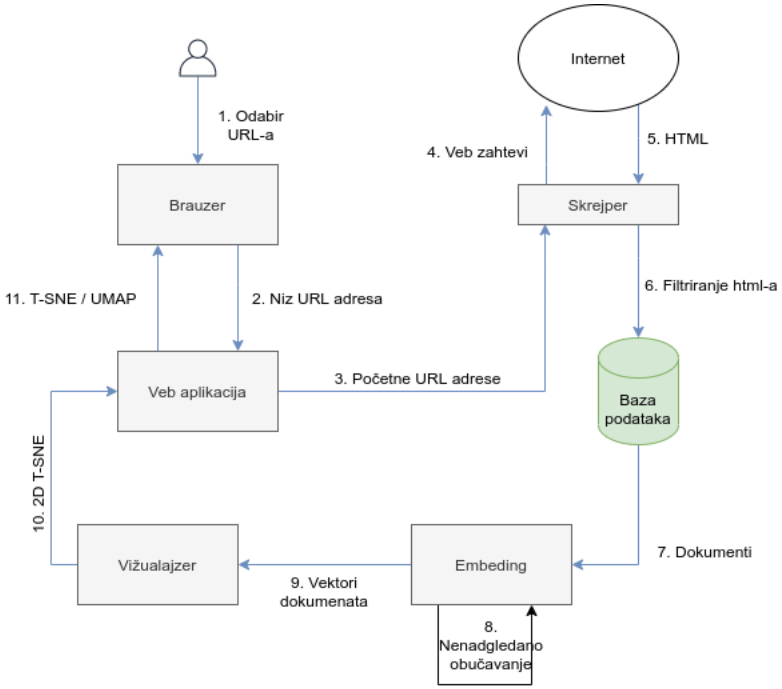
Pretraga prostih, jednostavnih činjenica prilično dobro funkcioniše sa postojećim pretraživačima. Ipak često je korisno i sagledavanje traženog pojma iz više uglova ili mišljenja kako bi se dobila objektivna slika traženog koncepta odnosno pojma. Trenutno ne postoji jednostavan način da se podaci isfiltriraju na ovaj način.

Čak i u slučaju uspešnog pronalaženja kvalitetnog dokumenta (veb stranice ili pdf fajla), ne postoji savšen način za njegovo dugotrajno skladištenje. Korisničko iskustvo skladištenja je gotovo nepostojeće za veb stranice.

Kako bi prethodni ciljevi bili postignuti, neophodno je imati jednostavan način skladištenja dokumenata i izvora, za koji ne bi bilo potrebno više od par sekundi. U suprotnom bi nastalo dodatno kognitivno opterećenje sortiranja i kategorizacije.

II. ANATOMIJA SISTEMA

U ovom delu će biti predstavljen pogled na sistem na visokom nivou. Zatim slede opisi sofverskih kompenenti od kojih se sistem sastoji. Gde je potrebno, opisi će biti upotpunjeni algoritmima i strukturama podataka iz kojih se sastoje.



Sl. 1. Blok šema Filternet arhitekture

A. Pregled arhitekture

Sistem je primarno napisan u programskom jeziku Pajton za Linuks operativni sistem. Ovaj programski jezik je odbaran zbog svoje fleksibilnosti, brzine dizajna prototipa, bogatog ekosistema i performansi (iako Pajton nije poznat po brzini, neretko su biblioteke za Pajton, tmo gde su performanse od krucijalnog značaja, napisane u jeziku nižeg nivoa).

Postupak rada sistema

U Filternet-u, korisnik započinje kompletan proces upotrebom ekstenzije pregledača na sledeći način (slika 1):

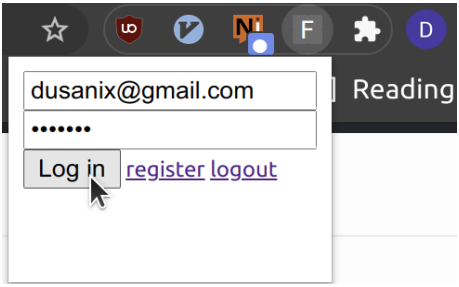
a) Nakon registracije na sajtu i uspešnog prijavljivanja ima mogućnost da sačuva stranicu (ili ceo portal) od interesa uz pomoć Filternet ekstenzije.

b) Veb aplikacija prihvata zahtev za indeksiranjem stranice i smešta ga u bazu podataka, kao i u FIFO red odakle će ga pročitati prvi slobodni skrejper.
c) Ukoliko stranica već nije procesirana ranije, skrejper će poštujući robots.txt ograničenja svakog veb sajta početi da preuzima sirove HTML podatke sa stranica u određenim vremenskim razmacima. HTML će biti parsiran po određenim pravilima gde će se izvući najkorisnije informacije. Ukoliko ne zadovoljava kriterijume stranica može biti odbačena.
d) Ekstraktovani podaci sa stranice će biti skladišteni u bazu podataka kao dokumenti. Istovremeno će se kreirati i vektori dokumenata (eng. embeddings) koji će takođe biti skladišteni u bazu podataka.
e) Za svakog korisnika se generiše mapa vektora spremna za prikazivanje.
f) Korisnik preko veb aplikacije može pristupiti mapi.

Brzina zahteva zavisi od broja zahtevanih veb stranica. Jednom zahtevane stranice se više ne skrejpuju ponovo za dva različita korisnika. Na nivou sisitema postoji baza već indeksiranih stranica.

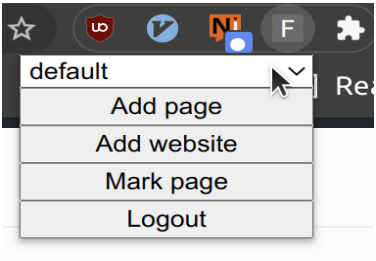
B. Brauzer ekstenzija

Filternet ekstenzija omogućava korisnicima jednostavno skladištenje veb stranica. Prvi korak u korišćenju ekstenzije je prijava na sistem, kako bi veb aplikacija mogla da zna koji korisnik je inicirao skrejpovanje. Klikom na ikonicu ekstenzije korisnik se može ulogovati radi kasnijeg dodavanja stranica (slika 2).



Sl. 2. Interfejs ekstenzije za prijavljivanje

Ekstenzija omogućava čuvanje veb stranice ili celog portala kao i odabir u koju grupu ih treba sačuvati (slika 3).



Sl. 3. Interfejs ekstenzije ulogovanog korisnika odakle je moguće dodavanje stranice ili celog sajta u grupu, odabir grupe kao i odjavljivanje.

Grupa predstavlja jednu tematsku celinu (npr: „magnetika“, „fizika“, „ruski pisci“, itd). Moguće je napraviti proizvoljno mnogo grupa. Svaka grupa odgovara jednoj mapi.

C. Filternet kroler

Pri svakodnevnom korišćenju interneta nailazimo na veb sajtove (npr. blogove) koji nam privuku pažnju ili su nam za nešto korisni. Razlozi mogu biti raznoliki. Uzmimo u razmatranje blog sa oko 200 stranica kvalitetnog sadržaja. Iako je moguće sačuvati ovaj blog u neki tekstualni fajl ili ga označiti (eng. to bookmark) u pregledaču, ovakav način skladištenja nije skalabilan. Sa povećanjem označenih adresa sve je teže pronaći traženu stranicu. Najveći problem kod ovog pristupa je podsećanje na stranicu. Iako je stranica sačuvana digitalno, da bi je uspešno povratili korisnici moraju da se :

a) Prisete da tu stranicu imaju (umni napor)
b) Pronađu je u postojećem sistemu čuvanja (može znatno da naraste i postane komplikovan za snalaženje)

Kroler komponenta se može posmatrati kao aktivni digitalni klon Filternet korisnika. Umesto korisnika će "pročitati" sve blog postove, zapamtiti ih i indeksirati za lakšu kasniju pretragu. Po potrebi se može i aktivirati u potrazi za novim blog postovima već sačuvanih autora i na ovaj način uštedeti vreme koje bi se utrošilo za ažuriranje postojećih stranica.

Ova komponenta je aktivni deo računarskog dela simbioze koji se periodično aktivira po potrebi korisnika, sa ciljem da olakša pristup odabranim informacijama. Korisnik ne mora da zapamti da je ikada sačuvao stranicu, ali će je lako prepoznati pri pretrazi sličnog pojma čak i nekoliko godina kasnije.

Uloga krolera

U prethodnom delu je opisana uloga ekstenzije u Filternet sistemu. Nakon što se zahtev prosledi veb aplikaciji preko ekstenzije i sačuva u bazi podataka, dalji posao preuzima kroler (eng. crawler).

Uloga krolera je da samostalno pregleda veb sajtove u potrazi za korisnim informacijama. Podešavanja ove komponente mogu odrediti način i obim potrage. Filternet podržava dva ograničena načina rada. Prvi je da se zahteva samo pojedinačna stranica. U ovom slučaju kroler nema mnogo posla i dovoljno je da za dati URL preuzme sirove HTML podatke samo jedne stranice. Drugi je da se zahteva domen. U tom slučaju korisnik zahteva krolovanje celog domena pri kojem kroler neće pratiti linkove ka eksternim adresama. Nakon što pronađe sve linkove na jednoj stranici i isfiltrira odlazne (linkove van trenutnog domena) kroler će podatke stranice prepustiti parseru.

D. Funkcije parsera

Komponenta koja se nalazi na čvorištu pohranjivanja i filtriranja je parser. Buduća poboljšanja pohranjivanja ili dodaci mogu se implementirati proširivanjem ovog programa. Nakon što kroler prikupi sirov sadržaj HTML stranice, parser preuzima ostatak posla.

Glavne funkcije parsera su:

1. Provera artikla
2. Generisanje dokument-vektora
3. Čuvanje u bazu podataka
4. Čuvanje u indeks

Provera da li je u pitanju članak

Pored tekstualnih stranica prikupljenog veb sajta (npr. bloga tj. blog posta)

postoje i funkcionalne stranice veb sajta kao što su: kontakt, o nama, kategorije, tagovi itd. Ove stranice, iako značajne za samo funkcionisanje veb sajta, nisu od velike koristi za Filternet. Iz ovog razloga je potrebno filtriranje sadržaja. Prvi korak parsera je da proveri da li je stranica koja se razmatra artikal, odnosno pretežno tekstualna stranica sa paragrafima. Filtriranje se vrši na osnovu heuristika. Poslednji korak je skidanje html tagova kako bi ostao čist tekst.

Generisanje dokument-vektora

Nakon prethodnog koraka gde su veb stranice filtrirane kao artikli i izbačeni su html tagovi sledi konverzija u vektor. Numerička predstava tekstualnih dokumenata ima različite namene. U zavisnosti od konteksta može se koristiti za pronalaženje dokumenata, filtriranje spama, ali i za tematsko modelovanje.

Konverzijom u niz brojeva odnosno vektor možemo porediti različite dokumente. Moguće ih je kreirati na dva načina:

- Vreća reči (eng. Bag of words) [12]
- Doc2Vec [13]

Čuvanje u bazu podataka i indeks

Nakon kreiranja vektora dokumenta potrebno ga je i sačuvati kako ne bi morao ponovo da se računa. Parser će u bazu podataka sačuvati osnovne podatke o stranici: URL, titl, md5 heš, dužinu stranice i vektor dokumenta. Na ovaj način se brzo može proveriti da li se sadržaj stranice promenio i da li je potrebno ponovo sračunati vektor dokumenta.

Poslednji korak je čuvanje dokumenta u indeks radi kasnije pretrage. U indeks se skladišti URL, titl, identifikacija reda baze podataka i čist tekst stranice. Nakon ove procedure skladištenje je kompletirano za kasnije pretraživanje.

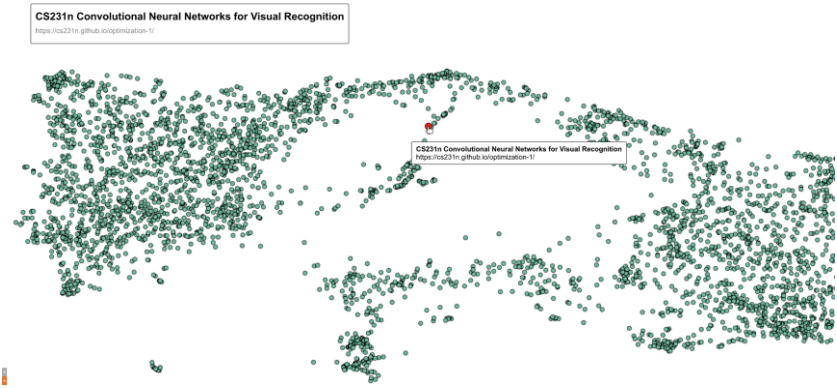
E. Interaktivna mapa

Iako je količina dostupnih podataka u konstantnom porastu, efikasni načini za njihovo prezentovanje se konstantno menjaju i razvijaju. Razvoj hardvera i sve veći protok interneta omogućili su modernim pregledačima (eng. browser)

da polako zamenjuju desktop aplikacije. Veb aplikacije rade na svim vodećim pregledačima (a samim tim i operativnim sistemima) i nemaju probleme portabilnosti koji otežavaju korišćenje desktop aplikacija. Osim toga, nemaju ni proceduru instalacije koja može biti naporna. Razvoj ekosistema vodećeg skript jezika za veb, JavaScript-a, omogućio je i napredne alate i biblioteke za razvoj kompleksnih grafičkih korisničkih programa na vebu.

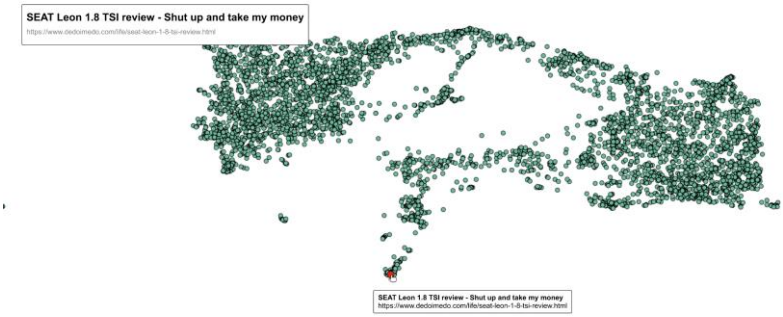
Filternet interaktivna mapa koja se generiše za svaku pojedinačnu grupu je napravljena u JavaScript jeziku. Veb aplikacija je odabrana za prezentovanje kao najfleksibilniji način za pravljenje kompleksnih dinamičnih prikaza podataka.

Interaktivna mapa je izabrana kao najprikladniji način za pregledanje potencijalno velikog broja veb stranica. Interaktivnost mape omogućava zumiranje samo onog dela mape za koji smo trenutno zainteresovani i sakriva nepotrebne delove mape čime se olakšava fokusiranje na određenu celinu.



Sl. 4. Vizuelni izgled Filternet mape.
Na slici je selektovana stranica indeksiranog sajta. U ovom prikazu su indeksirana dva veća bloga CodingHorror[7] i Dedoimedo [8] i nekoliko manjih sajtova. Ukupno oko 4500 stranica.

Svaka stranica predstavljena je krugom. Klikom na krug otvara se kartica sa naslovom stranice i linkom. Kombinacijom CTRL + klik se otvara stranica u novom tabu. Okretanjem točka na mišu moguće je odzumirati i zumirati interaktivnu mapu. Svaki krug je dimenzionalno redukovani vektor jedne stranice.



Sl. 5. Mini oblast.

Na slici 5. je selektovan članak o automobilu koji nema mnogo veze sa glavnom tematikom nijednog od indeksiranih blogova. Sve tačke oko ovog članka takođe predstavljaju, ili članak o konkretnom modelu automobila i njegovim karakteristikama, ili članak koji opisuje neki način transporta, ali takođe odudara od glavne tematike blogova, koja je softversko inženjerstvo i računarske nauke.

F. Pretraga

Kada se pomene reč pretraga u kontekstu interneta i računara, prva pomisao je uglavnom na generalnu pretragu kakvu nude popularni pretraživači na primer Google.com i Bing.com. Ovi pretraživači su nastali pre više od deceniju i nisu znatno promenili svoju funkcionalnost od tada. Od značajnijih promena ugrađena je pretraga mapa tj. geografskih referenci, vesti, slika i, u slučaju Google.com, pretraga imejl naloga, ukoliko je od iste kompanije, tj gmail.com. Iako su ovi dodaci korisni, i dalje se pretraga primarno odnosi na globalnu pretragu.

Sa sve većom količinom generisanih ili prikupljenih fajlova na ličnim računarima (npr. fotografije, tekst, prezentacije, knjige itd.) stvara se prirodna potreba za lakšim upravljanjem, a time i pronalaženjem ovih fajlova. Većina operativnih sistema nudi relativno sporu pretragu po nazivima fajlova. Ako bismo na trenutak zamislili da i veliki pretraživači rade po istom principu

korišćenje interneta bi bilo znatno teže. Količina podataka na veb stranici je znatno veća nego string koji opisuje tag sa naslovom stranice (eng. title tag).

Čuvanje veb stranica od interesa samo pomoću operativnog sistema personalnog računara je nezgrapno. Iako bi pretraga mogla da se izvede pomoću široko dostupnih alata (npr. grep na linuksu), ovakva upotreba zahteva naprednije poznavanje računara i korišćenje terminala. Transakciona cena skladištenja je relativno velika. Fajl sistem operativnog sistema nije pogodan za ovaj zadatak. Koraci preuzimanja veb stranice su:

1. Manuelno čuvanje stranice iz brauzera
2. Odabir foldera u kom će biti sačuvana stranica
    a. Ovde je poželjno dodatno kategorisanje foldera, jer u slučaju velikog broja stranica kasnije pronalaženje može biti problematično

Zbog prethodno navedenih razloga, ovaj način skladištenja je prilično opterećujući, pogotovo za veći broj stranica (npr. skladištenje kompletnog jednog bloga).

Pretraživanje ličnih podataka je donekle otežano načinom na koji veliki pretraživači zarađuju. Kako bi zaradili od svojih usluga postavljaju reklame u rezultate pretrage. Ukoliko reklame nisu relevantne korisniku, verovatnoća da klikne na neku od njih znatno opada. Sa ciljem da povećaju profit, kompanije neretko analiziraju sve podatke i metapodatke korisnika kako bi povećali verovatnoću kliktanja, tj zarade. Sa sve većim digitalnim otiskom je sve lakše napraviti precizan profil korisnika. Čak i sa svesnim trudom od strane korisnika da smanji svoj digitalni otisak, mnogi zaključci se mogu izvući samo od obrazaca ponašanja i akcija koje preduzima na internetu. Finansijska motivacija kompanije nije u potpunosti usaglašena sa pravom na privatnost korisnika i otvara prostor za inovacije.

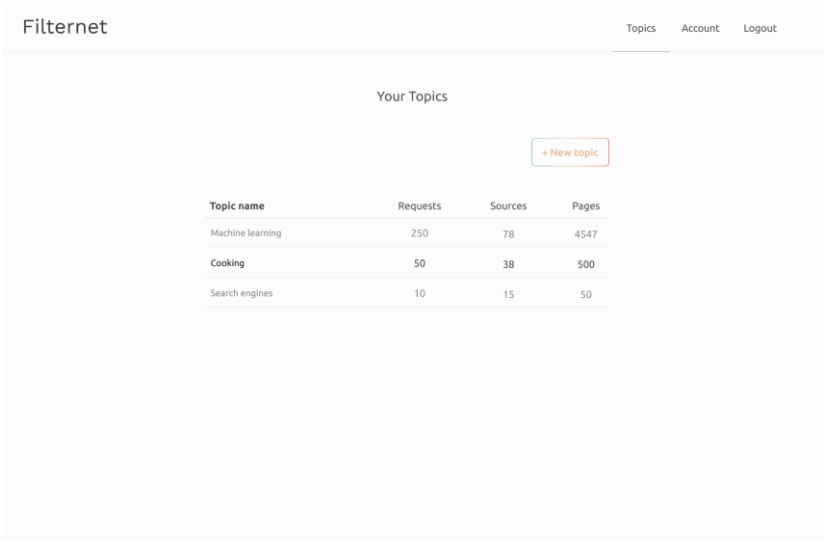
Osim kompanije koja nudi usluge, moguće je da i treća lica dođu do podataka kao u slučaju kompanije Kembridž-Analitika. Ukratko, u ovom slučaju [10] je eksterna kompanija skrejpovala podatke sa Fejsbuk profila stotine hiljada ljudi i precizno ih podelila u hiljade interesnih grupa. Zatim su unajmili kreativne timove i psihologe koji bi kreirali digitalne sadržaje (reklame, zanimljive fejsbuk stranice, memeove - slike duhovitog karaktera itd.) za svaku interesnu grupu, sa ciljem da efektivno utiču na korisnike Fejsbuka.

Prethodno navedeni primeri ilustruju potrebu za skladištenjem podataka od interesa, ali i za pretragom tih podataka koja poštuje privatnost. Filternet pretraga je osmišljena kao "full-text" pretraga, koja je kreirana sa ciljem da pretraži podatke iz računarskog dela simbioze čovek-računar poštujući privatnost korisnika.

G. Veb aplikacija za upravljanje

Dopunjavanje mape uz pomoć brauzer ekstenzije je proces u kom se mapa proširuje novim sadržajima. Ipak, naknadno je potrebno upravljati mapama, što podrazumeva:

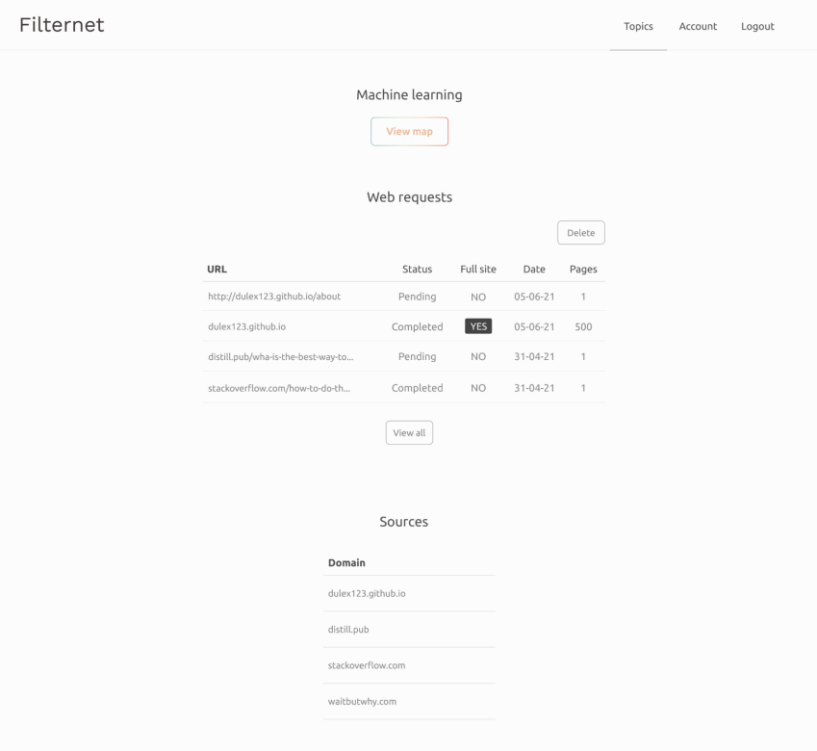
- Dodavanje novih mapa (odnosno novih tematskih celina)
- Brisanje mapa
- Brisanje već dodatih sadržaja
- Premeštanje sadržaja



Sl. 6. Glavni panel veb aplikacije.

Stranica koja se prva otvara prilikom prijavljivanja na veb aplikaciju je glavni panel (eng. Dashboard). Tabela na sredini prikazuje naziv tematske celine i informacije o broju zahteva, različitih sajtova od kojih se sastoji i ukupnog

broja stranica koji čine tematsku celinu. Osim odabira neke od postojećih, moguće je i dodati novu.



Sl. 7. Prikaz stranice za upravljanje jednom tematskom celinom

Prva tabela pokazuje od kojih zahteva se sastoji trenutno posmatrana tematska celina i dodatnih informacija. URL označava početni link, status - da li je već indeksiran link, "Full site" - da li je zahtevano indeksiranje celog sajta ili samo jednog linka, datum zahteva i ukupan broj strana. Ispod prve tabele se nalazi lista izvora i naznačeni su domeni koji sačinjavaju temu. Ovakvim izlistavanjem je olakšano uklanjanje kompletnih domena (ukoliko je bilo više zahteva za isti domen).

III. ZAKLJUČAK

Razvoj veštačke inteligencije je spor i dugotrajan proces. Ipak, daleko je brži nego evolutivni razvoj ljudskog mozga. Računari oko nas su neizostavni deo

svakodnevnice sa trendom sve većeg zbližavanja. Startap kompanije već uveliko rade na direktnom iščitavanju moždanih signala radi većeg protoka informacija između računara i čoveka. Fizička simbioza, gde bi računar bio implant u ljudskom telu, više nije tako daleko.

Računari olakšavaju kognitivne napore ljudskog mozga od svog nastanka. Na ovakav način se može uzeti u razmatranje i trivijalni (za računar) kalendarski podsetnik. Ljudska memorija je nepouzdana u poređenju sa računarskom. Usled različitih emocija, stresa ili euforije možemo zaboraviti bitne informacije u bitnim trenucima. Rasterećivanjem uma uz pomoć računara možemo osloboditi umni kapacitet za razmišljanje. Istraživanje može biti naporan zadatak sa mnogo nepoznatih. Filternet je zamišljen kao digitalni pomoćnik prilikom analize novih oblasti. U ovom radu je opisan princip rada sistema za tematsko arhiviranje i pretraživanje dokumenata sa fokusom na veb stranice. Razvijan je sa idejom istraživača kao primarnog korisnika.

LITERATURA

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville “Deep Learning”, MIT Press, 2016
[2] Laurens van der Maaten, Geoffrey Hinton, “Visualizing Data using t-SNE", Journal of Machine Learning Research, 2008
[3] OpenSyllabus project, http://opensyllabus.org/
[4] Google Maps, http://www.google.com/maps/
[5] OpenSyllabus Galaxy, http://galaxy.opensyllabus.org/
[6] Thomas Cormen, Charles Leiserson, Ronald Rivest, Clifford Stein, MIT Press, 2009
[7] CodingHorror blog, https://blog.codinghorror.com/
[8] Dedoimedo blog, https://www.dedoimedo.com/
[9] D3.js project, https://d3js.org
[10] *Cambridge Analytica data scandal, https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal*
[11] *Apache Lucene, https://lucene.apache.org/*
[12] Tomas Mikolov, Efficient Estimation of Word Representations in Vector Space, https://arxiv.org/abs/1301.3781
[13] Quoc Le, Distributed Representations of Sentences and Documents, https://cs.stanford.edu/~quocle/paragraph_vector.pdf
[14] ObservableHQ D3.js Gallery, https://observablehq.com/@d3/gallery

ABSTRACT

Personal knowledge management is a process that the user uses to efficiently collect, classify, store, search, review, and share knowledge during day-to-day activities. With the growing number of digital content created, this process is

becoming increasingly difficult. This paper presents a system for thematic archiving and data retrieval, with a focus on web pages, which facilitates the management of personal knowledge. The purpose of the presented system is to help in research in which it is necessary to adopt a larger number of interrelated concepts. The main processes of the system are scraping web pages and embedding, which provides an interactive map to display thematic areas. The paper conceptually presents the components of the system and analyzes an example of a specific implementation. The result is a practical implementation for use and further study.

Filternet: System for topical filtering, browsing and archiving of data

Dušan Josipović, Dušan Vujošević