Ciobănița Andreea Sorana

Filtrarea conținutului la nivel de client pe baza interesului utilizatorilor

Cuprins

[1. Introducere 3](#_Toc72697176)

[1.1. Prezentare generală 3](#_Toc72697177)

[1.2. Motivație 3](#_Toc72697178)

[2. Fundamente Teoretice 5](#_Toc72697179)

[2.1. Filtrarea conținutului web la nivel de client 5](#_Toc72697180)

[2.2. Soluții existente 5](#_Toc72697181)

[2.3. Rezultate teoretice 6](#_Toc72697182)

[3. Soluția propusă 7](#_Toc72697183)

[3.1. Utilizare 7](#_Toc72697184)

[3.2. Detalii de implementare 7](#_Toc72697185)

[4. Evaluarea rezultatelor 8](#_Toc72697186)

[5. Concluzii 9](#_Toc72697187)

[Bibliografie 10](#_Toc72697188)

# Introducere

## Prezentare generală

Filtrarea conținutului web este reprezentată de tehnicile de blocare a afișării diferitor elemente din paginile web, în funcție de diferite criterii.

În primul rând, necesitatea unei filtrări la nivelul conținutului paginilor web a apărut odată cu facilitarea accesului la internet, deoarece persoane de toate vârstele au învățat să îl folosească în viața de zi cu zi. Din această cauză, a apărut nevoia de implementare a unor metode ce să poată oferi un control parental, ajutând astfel părinții să își protejeze copiii împotriva conținutului nepotrivit ce poate fi foarte ușor accesibil pe internet.

O altă utilitate a filtrării este folosirea acesteia în instituții și companii, pentru a limita site-urile la care un utilizator să aibă acces în cadrul rețelei interne de internet. Acest lucru ajută, de exemplu, în rețelele folosite în școli pentru a împiedica elevii să acceseze site-uri cu conținut nepotrivit sau diferite platforme care ar putea afecta calculatoarele instituției prin instalarea unor viruși. De asemenea, companiile pot folosi un sistem de filtrare pentru a împiedica angajații să petreacă prea mult timp navigând pe diferite pagini non-productive, astfel fiind posibilă o creștere a eficienței și a productivității în timpul programului de lucru.

De asemenea, o altă utilitate importantă este posibilitatea de a filtra conținutul pe baza intereselor utilizatorilor. Datorită creșterii foarte mari din ultimii ani a internetului, au apărut foarte multe informații ușor accesibile, însă uneori când căutăm anumite lucruri de care suntem interesați primim și foarte multe reclame care de cele mai multe ori acoperă conținutul paginilor. Uneori ceea ce primim ca și reclamă ar putea fi ceva de interes pentru noi, mai ales că în ultima perioadă și sistemele de recomandare și afișare a reclamelor s-au dezvoltat, fiind capabile să prezinte cât mai mult lucruri din aria de interes a utilizatorilor, însă de cele mai multe ori când navigăm pe internet am dori să ne putem concentra atenția pe lucrurile pe care le căutăm în momentul respectiv. Din acest motiv s-au dezvoltat diferite aplicații software de tip ad-blocker pentru a ascunde cât mai eficient reclamele nedorite ce apar în timpul navigării pe internet.

## Motivație

Deși ad-blocker-ele au adus un mare avantaj în navigarea pe internet, acestea nu reușesc întotdeauna să ascundă toate reclamele care apar. În plus, filtrarea conținutului pe baza intereselor utilizatorilor poate fi extinsă să însemne mai mult decât ascunderea unor reclame. Există situații în care ne putem dori să ascundem dintr-o anumită pagină și altfel de elemente, care nu au cum să fie incluse în cele ce sunt filtrate de un ad-blocker.

De exemplu, daca navigăm pe un site de știri și suntem interesați de cele dintr-un anumit domeniu, ar fi mai ușor dacă am avea posibilitatea ca în momentul în care găsim o știre din afara domeniului nostru de interes să avem opțiunea de a o ascunde din pagină. Astfel, în final putem vizualiza mult mai ușor ceea ce ne interesează de fapt, fără să ne fie distrasă atenția cu alte lucruri. Un alt exemplu din această categorie ar putea fi diferite animații care se mișcă în continuu undeva pe pagina pe care o vizualizăm și am dori să putem să le ascundem.

În acest sens, propunem o extensie Google Chrome care să ne ofere posibilitatea de a selecta și apoi a ascunde elementele pe care nu dorim să le vedem afișate în paginile pe care navigăm. Odată instalată în browser, extensia va adăuga o opțiune numită ”*Remove Element*” în meniul afișat în momentul în care dăm click dreapta pe un element din pagină, iar aceasta va ascunde elementul selectat. Mai târziu, în fereastra extensiei va fi afișată o listă cu elementele care au fost ascunse și va exista posibilitatea ca acestea să fie scoase din listă și astfel să fie afișate din nou în pagină la locul lor. De asemenea, ne propunem să folosim anumite keywords din textul elementelor pe care le ascunde un utilizator și după câteva utilizări să putem oferi sugestii în legătură cu posibile elemente ce nu par a fi de interes pentru acesta și ar putea fi ascunse din paginile pe care navighează.

# Fundamente Teoretice

## Filtrarea conținutului web la nivel de client

Un program software pentru filtrarea conținutului web este un program construit pentru a controla ceea ce un utilizator poate vedea în pagini atunci când navighează pe Internet, sau pentru blocarea în întregime a anumitor URL-uri. Există două metode de a filtra conținutul web, fie se vor folosi liste predefinite cu site-uri unde utilizatorului îi este permis să navigheze, fie liste cu elemente care nu sunt permise. În principal, programele de acest tip sunt integrate în antiviruși, instalate ca și programe de sine stătătoare care sunt rulate pe computere sau extensii pentru browser-ul web.

Filtrarea conținutului direct la nivelul clientului web, reprezentat de browser, este de cele mai multe ori preferată de utilizatori, deoarece este mai ușor de folosit și nu implică existența altor programe adiționale. Printre cele mai frecvente utilități ale unei extensii pentru browser care se ocupă de filtrarea conținutului web se numără blocarea reclamelor și restricționarea folosită în scopul controlului parental.

Un dezavantaj al extensiilor este faptul că acestea sunt configurate pentru un browser anume, iar în cazul în care utilizatorul va folosi altul, filtrarea nu va mai avea loc până nu vor fi adăugate extensiile necesare și în browser-ul nou. De asemenea, de multe ori extensiile sunt disponibile doar pentru un anumit browser, nu pentru toate. Din aceste motive uneori sunt preferate programele independente care să fie capabile să ruleze și să realizeze filtrarea indiferent de browser.

## Soluții existente

În lucrarea ”FilterPlus: A real-time content filtering extension for Google Chrome” [1] autorii prezintă extensia FilterPlus. Aceasta oferă o interfață simplă prin care utilizatorii pot filtra imagini, ferestrele de tip popup, codul JavaScript, Cookie-urile și notificările pe paginile web pe care navighează. Permițând blocarea imaginilor, a codului JavaScript și a popup-urilor este asigurat un aspect mai curat al paginilor și în același timp este crescută securitatea, deoarece numeroase programe de tip malware sunt instalate prin intermediul codului rulat fără știrea utilizatorilor. Pentru cookies sunt disponibile trei opțiuni: Allow, Session Only și Block, astfel oferind control utilizatorului asupra a ceea ce dorește sau nu să fie memorat, iar pentru notificări este posibil ca acestea să fie acceptate, blocate, sau acceptate doar după ce extensia va întreba înainte pentru fiecare în parte.

Thomas Chen și Victoria Wang prezintă în ”Web Filtering and Censoring” [2] aplicația Green Dam, care este o aplicație desktop ce filtrează conținutul web. Folosește un blacklist definit de utilizator pentru a filtra mai întâi URL-urile unde accesul nu este permis, apoi folosește analiza de text pentru a căuta și filtra posibile obscenități sau conținut sensibil politic. De asemenea, implementează și o metodă de filtrare a imaginilor pornografice.

” Dynamic Web Content Filtering based on User’s Knowledge” [3] descrie un sistem de filtrare pe baza conținutului, ca o soluție pentru problema filtrării după URL, deoarece aceasta nu este eficientă pentru site-urile web necunoscute încă și o listă completă de URL-uri ce nu ar trebui permise este imposibil de obținut. Pentru implementarea soluției se folosește metoda de achiziționare a cunoștințelor cu reguli multiple de clasificare - *Multiple Classification Ripple-Down Rules (MCRDR),* care permite unui expert în domeniu să mențină baza de cunoștințe fără să fie nevoie de ajutorul inginerilor. Sistemul este antrenat pe un anumit set pagini web pentru a învăța care sunt cele ce trebuie să fie blocate. Ca și exemplu, sistemul prezentat învață să clasifice un set de 283 site-uri web în pagini care au sau nu legătură cu jocurile de noroc.

## Rezultate teoretice

Prima soluție prezentată, FilterPlus, reușește cu succes să crească securitatea utilizatorilor oferind posibilitatea de blocare a codului JavaScript și a ferestrelor de tip popup și, în același timp, oferă și o filtrare a conținutului permițând ascunderea imaginilor sau a notificărilor. Preferințele utilizatorilor se reflectă însă doar în alegerea de a permite sau nu elementele pe care extensia le poate bloca, însă nu oferă posibilitatea de a alege altceva, de a filtra conținutul mai mult decât atât, pe baza altor reguli.

Aplicația Green blochează în mod eficient toate URL-urile aflate pe lista neagră, iar sistemul de filtrare pe baza analizei de text este destul de precis. Pe de altă parte, sistemul de blocare a imaginilor pornografice nu este la fel de eficient, deoarece acesta a clasificat ca și conținut nepotrivit și imagini greșite, precum imagini cu o pisică sau o farfurie cu friptură. De asemenea, filtrarea pe baza altor preferințe ale utilizatorilor nu este posibilă nici în cadrul aplicației Green.

Ultima soluție descrisă este un sistem destul de eficient de filtrare pe baza preferințelor utilizatorului, întrucât acesta poate alege ce fel de pagini dorește să fie blocate în funcție de conținutul lor. Un dezavantaj ar fi faptul că este necesar un set de câteva pagini pentru învățare, pentru ca sistemul să fie capabil mai apoi să clasifice site-urile nou accesate.

Toate cele trei soluții prezentate au atât avantaje cât și dezavantaje și se potrivesc pentru anumiți utilizatori, în funcție de nevoile lor. În timp ce primele două nu permit o personalizare a conținutului ce se dorește a fi blocat, ultima oferă această posibilitate. Pe de altă parte, chiar dacă a treia soluție implementează filtrarea pe baza preferințelor utilizatorului, aceasta blochează întreaga pagină web, nefiind posibilă ascunderea doar pentru elementele ce nu sunt de interes.

# Soluția propusă

Soluția propusă este o extensie pentru browser-ul *Google Chrome*, care va permite utilizatorilor să aleagă elementele din pagini web de care nu sunt interesați și să le ascundă. Spre deosebire de celelalte soluții descrise, extensia implementată nu are anumite filtre predefinite pentru ce fel de elemente să fie blocate, precum imaginile sau textul având conținut nepotrivit, ci oferă posibilitatea utilizatorului să aleagă orice element din pagină de care nu este interesat și să îl ascundă. De asemenea, extensia nu va bloca întreaga pagină web, ci doar va ascunde elementele selectate de utilizator care nu sunt de interes.

## Utilizare

Extensia adaugă o opțiune nouă în meniul ce apare atunci când utilizatorul dă click dreapta pe un anumit element dintr-o pagină web, numită ”*Remove Element*”. Cu ajutorul acestei opțiuni, elementul ales va fi ascuns din pagină iar informațiile despre acesta vor apărea într-un tabel în popup-ul extensiei. Pentru fiecare element va fi disponibilă de asemenea opțiunea de a-l șterge din listă, astfel acesta va apărea din nou în pagină în locul său inițial. Elementele ce au fost ascunse vor fi memorate, astfel încât la o nouă accesare a aceleași pagini utilizatorul să nu fie nevoit să le ascundă din nou, acest lucru având loc în mod automat.

Mai mult decât atât, în momentul ascunderii mai multor elemente ce conțin text într-o pagină web, aplicația va încerca să găsească niște similarități, alcătuind o listă de cuvinte întâlnite mai des în conținutul de care utilizatorul nu este interesat. Pe baza acestor keywords extensia oferă opțiunea de a sugera alte elemente ce ar putea fi eliminate, iar utilizatorul poate alege dacă vrea să ascundă elementele sugerate.

## Detalii de implementare

Interfața aplicației, și anume fereastra de tip popup în care vor fi afișate elementele ascunse de către utilizator a fost implementată cu ajutorul framework-ului *React*. Fiind vorba despre filtrarea la nivel de client, funcționalitățile aplicației sunt implementate exclusiv în *JavaScript*, în cadrul extensiei, fără a implica apeluri către alte surse externe. Pentru memorarea elementelor a fost folosit state-ul global al aplicației cu ajutorul *React Hooks*.

Pentru găsirea de similarități în text, alcătuirea listei de cuvinte cheie și sugerarea posibilelor elemente pe care utilizatorul ar putea dori să le ascundă s-au folosit librăriile JavaScript *string-similarity* și *keyword-extractor*.

# Evaluarea rezultatelor

Pentru testarea extensiei s-a folosit site-ul Facultății de Matematică și Informatica, Universitatea Babeș-Bolyai, <http://www.cs.ubbcluj.ro/>.

Funcționalitatea de ascundere a elementelor selectate, întrucât este una destul de simplă, funcționează bine de fiecare dată. Așadar, ne vom axa pe evaluarea funcționalității de sugerare a elementelor similare cu cele ascunse ulterior de utilizator.

Pentru testare s-a încercat ascunderea anunțurilor cu informații despre admiterea la studiile de nivel licență și s-a verificat dacă elementele sugerate anterior fac parte din aceeași categorie. Primul anunț care a fost ascuns conținea informații despre organizarea concursului de admitere în luna aprilie 2021. În urma ascunderii acestui articol, a început formarea listei de cuvinte cheie, însă fiind doar primul element ascuns au rezultat 43 de cuvinte cu posibilitatea de a fi cuvinte cheie. Așadar, aplicația nu va genera sugestii după ce un singur element a fost ascuns. Al doilea articol ales conținea informații despre calendarul concursului de admitere, iar în urma extragerii de cuvinte cheie din acesta s-au obținut 41 de variante. Făcând însă intersecția între cele două liste de cuvinte, a rămas un set de 8 keywords care să fie utilizat pentru căutarea altor elemente ce ar putea fi ascunse.

Pe baza acestei liste de cuvinte, aplicația a verificat alte anunțuri pentru a sugera unele care e posibil să nu fie de interes. Un prim articol care a fost sugerat a fost cel cu rezultatele concursului de admitere la nivel licență, o sugestie conform așteptărilor. Întrucât printre cuvintele cheie se aflau ”calendar” și ”admitere”, a fost de asemenea sugerat și un anunț despre calendarul de admire la nivel de master. Chiar dacă acest anunț nu este legat de studiile la nivel de licență, nu putem spune ca sugestia a fost întocmai greșită, deoarece articolul conține o parte dintre cuvintele cheie.

Un dezavantaj al site-ului ales este faptul ca fiecare anunț este afișat cu titlu și un prim paragraf, urmând apoi un buton cu textul ”Vezi restul anunțului” pentru a accesa întregul articol. Din cauză că atunci când au fost ascunse articolele pentru testare acest buton a fost găsit în ambele, extensia a sugerat de asemenea ascunderea celorlalte butoane de acest tip.

Așadar, funcția de sugerare a unor elemente asemănătoare cu altele ascunse anterior de utilizatori pe baza unor cuvinte cheie este destul de precisă, însă din cauza faptului că se folosesc doar aceste cuvinte cheie rezultatul este unul foarte general. În acest fel, sugestiile nu vor putea fi cuprinse într-un anumit domeniu care nu este de interes, ci vor reprezenta orice alte elemente care conțin în textul lor o parte dintre cuvinte.

Extensia ar putea fi îmbunătățită schimbând modul în care sunt extrase cuvintele cheie, alegând grupuri de cuvinte în schimbul celor individuale, pentru a crește precizia. De asemenea, ar fi bine să fie luată în considerare și frecvența de apariție a cuvintelor în texte, iar daca un keyword nu apare de un anumit număr minim de ori să nu fie luat în considerare mai departe.

# Concluzii

Odată cu dezvoltarea imensă a Internetului din ultimii ani și cu faptul că foarte multe persoane au început să îl folosească pentru activitățile de zi cu zi, a apărut și necesitatea unei filtrări a conținutului web. În afară de motivele de securitate, control parental sau blocare a reclamelor ce acoperă conținutul paginilor, datorită multitudinii de informații ce pot fi prezente pe un singur site, a apărut și nevoia de a filtra elemente care pur și simplu nu fac parte din aria de interes a utilizatorului.

Aplicația propusă este o extensie pentru browserul Google Chrome, care permite utilizatorilor să ascundă din paginile web pe care navighează orice element doresc. Mai mult decât atât, după ce au fost ascunse câteva elemente ce conțin text, extensia poate sugera alte elemente pe care le găsește ca fiind similare.

Funcționalitatea de ascundere a elementelor selectate de utilizator are o precizie foarte bună, eliminând cu succes din pagini orice elemente alege acesta. Partea de sugerarea a elementelor asemănătoare ar putea fi îmbunătățită. Aceasta se bazează pe căutarea de cuvinte cheie în textele ascunse anterior, iar din această cauză aria de căutare este destul de largă și uneori pot fi sugerate și elemente care aparent nu au legătură cu domeniul celor anterioare, având în comun doar câteva dintre cuvintele cheie.

# Bibliografie

|  |  |
| --- | --- |
| [1] | B. Babu și M. Kumar, „FilterPlus: A real-time content filtering extension for Google Chrome,” 2014. |
| [2] | T. M. Chen și V. Wang, „Web Filtering and Censoring,” *IEEE Computer,* 2010. |
| [3] | N. Churcharoenkrung, Y. S. Kim și B. H. Kang, „Dynamic web content filtering based on user's knowledge,” *Proceedings of the International Conference on Information Technology: Coding and Computing.* |
| [4] | E. Bertino, E. Ferrari și A. Perego, „A general framework for web content filtering,” *World Wide Web,* 2010. |
| [5] | W. Ming, L. Rossi, Y. Li și J. Kuo, „Automatic Web Filtering Approach Based on Multimodal Content Information,” *Internet Multimedia Management System II,* 2001. |
| [6] | C. Ding, C.-H. Chi, J. Deng și C.-L. Dong, „Centralized content-based web filtering and blocking: How far can it go?”. |
| [7] | Y. Lai, Q. Ma, Z. Yang și J. Liu, „Framework of Web Content Filtering for IPv6”. |
| [8] | P. Y. Lee, S. C. Hui și A. C. M. Fong, „Neural Networks for web content filtering”. |
| [9] | D. D. Nguyen, M. Erdmann, T. Takeyoshi și G. Hattori, „Training multiple support vector machines for personalized web content filters,” *IEICE Trans. Information & Systems,* 2013. |
| [10] | N. Narwal și S. K. Sharma, „Web informative content identification and filtering using machine learning technique,” *Data Analysis Techniques and Strategies,* 2016. |
| [11] | A. C. M. Fong, S. C. Hui și P. Y. Lee, „XFighter: An intelligent web content filtering system,” *Kybernetes,* 2016. |
| [12] | K. S. Kuppusamy și G. Aghila, „A personalized web page content filtering model based on segmentation,” *International Journal of Information Sciences and Techniques,* 2012. |
| [13] | M. Chau și H. Chen, „A machine learning approach to web page filtering using content and structure analysis,” 2007. |
| [14] | T. Sandholm și H. Ung, „Real-time, location-aware collaborative filtering of web content,” 2011. |
| [15] | M. Hammami, Y. Chahir și L. Chen, „WebGuard: a web filtering engine combining textual, structural and visual content-based analysis,” *IEEE Transactions on Knowledge and Data Engineering,* 2006. |
| [16] | P. Y. Lee, S. C. Hui și A. Fong, „A structural and content-based analysis for Web filtering,” *Internet Research: Electronic Networking Applications and Policy,* 2003. |
| [17] | E. Bertino, E. Ferrari, A. Perego și G. P. Zarri, „An Integrated Approach to Rating and Filtering Web Content,” 2005. |
| [18] | G. Sartor și A. Loreggia, „The impact of algorithms for online content filtering or moderation,” 2020. |
| [19] | E. Bertino, E. Ferrari și A. Perego, „Content-based filtering of web documents: the MaX system and the EUFORBIA project,” 2003. |
| [20] | L. Yan și Y. Shenghua, „Content filtering research based on web community structure,” *International Conference of Mechatronic Sciences, Electric Engineering and Computer (MEC),* 2013. |
| [21] | S.-J. Ko, „Predicting Typical User Preferences Using Entropy in Content Based Collaborative Filtering System,” 2004. |
| [22] | W. Paireekreng, „Mobile content recommendation system for re-visiting user using content-based filtering and client-side user profile,” *International Conference of Machine Learning and Cybernetics,* 2013. |