Universitatea Babeş - Bolyai
Facultatea de Matematică şi Informatică

Relevanţa conţinutului web şi a comportamentului utilizatorilor în analiza traficului

Proiect de cercetare

Student doctorand: Diana-Florina HALIŢĂ
Coordonator ştiinţific: Prof. Dr. Florian Mircea BOIAN

Iunie 2015

Cuprins

Introducere

Navigarea pe Internet a devenit un aspect esenţial din viaţa de zi cu zi a fiecăruia dintre noi. Totodată aceasta începe să aibă din ce în ce mai mult o dimensiune socială, devenind astfel un mecanism eficace prin intermediul căruia se pot dobândi cunoştinţe. Dezvoltarea rapidă a World Wide Web, precum şi varietatea resurselor care sunt disponibile în Internet, impun necesitatea existenţei unor instrumente care pot extrage informaţii importante din resursele existente sau care pot duce la îmbunătăţirea calităţii experienţelor utilizatorilor în timpul explorării acestora.

Astfel de instrumente joacă un rol important atât pentru administratorii site-urilor web, cat şi pentru utilizatorii obişnuiţi. Din perspectiva unor experţi interesul major ar fi capturarea şi analizarea traficului web, cu scopul de a deservi interesul utilizatorilor.

În acest context, lucrarea de faţă se concentrează pe prezentarea a două abordări diferite. Prima abordare se referă la descoperirea de informaţii importante disponibile pe web. Cea de-a doua se referă la urmărirea şi analizarea comportamentului utilizatorilor şi a modalităţilor în care acesta este influenţat de conţinutul care se regăseşte online. De ce aceste două abordări?

Analiza conţinutului, ca tehnică de cercetare, este obiectivă, sistematică şi poate defini documentul atât din punct de vedere cantitativ cât şi din punct de vedere calitativ.

Privind prima abordare urmărită în cercetarea noastră, luând în considerare analiza unui document care este interesant sau util, se poate pune problema găsirii de documente similare care să prezinte un interes la fel de mare pentru studiul considerat.

În literatura de specialitate, tema abordată pentru realizarea acestei lucrări se află la confluenţa mai multor clase: analiza legăturilor dintre documente şi analiza consecinţelor ce pot apărea în urma acestor legături.

Privitor la cea de-a doua abordare, scopul acestei lucrări este acela de a găsi noi modalităţi de analizare a comportamentului utilizatorilor unui site, astfel încât acesta să poată fi adaptat permanent nevoilor acestora (atât din punct de vedere al design-ului, cât şi din punct de vedere al tipului de conţinut prezentat). Majoritatea cercetătorilor propun utilizarea tehnicilor din *Web Usage Mining* pentru a putea studia astfel de comportamente.

Motivaţia alegerii temei

Alegerea temei a fost influenţată de câteva probleme concrete cu care autorul acestei lucrări împreună cu membrii grupurilor de cercetare din care face parte s-a confruntat în ultimul timp.

Obiective:

- migrarea transparentă a unui site web atât din perspectiva utilizatorilor, cât şi din perspectiva motoarelor de căutare, asistarea motoarelor de căutare în redirecţionarea corectă a vizitatorilor, precum şi redirecţionarea consecventă a vizitatorilor care ajung pe site, prin intermediul terţilor *referreri*;
- estimarea ratei de respingere pentru terţe site-uri cu posibile aplicaţii în ierarhizarea site-urilor web, precum şi a *spam linking*-ului şi a *spamdexing*-ului;
- identificarea unei măsuri numerice de cuantificare a nivelului de *scrapering* a unui site precum şi vizualizarea caracteristicilor unor astfel de site-uri;
- identificarea unor traiectorii in diverse seturi de date asociate unor site-uri cu scopul de a folosi informaţiile identificate în reorganizarea acestora din punct de vedere structural sau vizual;
- detectarea comportamentului utilizatorilor într-o platforma de *e-learning* folosind Analiza Conceptuală Formală;
- identificarea de structuri triadice în logurile web ale platformei de *e-learning* şi utilizarea acestora în determinarea dinamicii temporale a utilizatorilor, folosind Analiza Conceptuală Formală Triadică;
- identificarea *trendsetteri*lor care creează diverse traiectorii prin site şi a grupurilor de utilizatori care urmează traiectorii similare, folosind Analiza Conceptuală Triadică şi Analiza Conceptuală Formală Temporală;
- urmărirea interesului utilizatorilor faţa de informaţia prezentată pe platformă, considerând scopul pentru care aceasta a fost construită.

Structura

Lucrarea de faţă este structurată în 2 capitole, după cum urmează.

Capitolul I - Impactul similarităţii web asupra traficului prezintă prin intermediul celor trei secţiuni principale, modalitatea în care similaritatea documentelor regăsite online îsi pune amprenta asupra traficului web.

Secţiunea I - Migrarea transparentă a unui site web între două sisteme de management de conţinut prezintă paşii care ar trebui urmaţi astfel înât să se poată face migrarea transparentă a unui site web de la un Sistem de Management bazat pe Conţinut (CMS) la altul. Astfel, interesul cade asupra a ceea ce se poate face în privinţa vizitatorilor site-ului care ar putea fi direcţionaţi greşit de motoarele de căutare sau de terţi *referreri* la *URL*-uri vechi, inexistente.

Secţiunea II - Estimarea *bounce-rate*-ului pentru terţe site-uri cu posibile aplicaţii în ierarhizarea site-urilor, precum şi a *spam linking*-ului analizează posibilitatea existenţei unei legături între similaritatea documentelor web şi a ratei de respingere generate de *link*-urile dintre aceste documente şi identificarea *link*-urilor abuzive prin compararea conţinuturilor aflate în pagina care conţine *link*-ul şi pagina spre care duce *link*-ul.

Secţiunea III - Concluzii şi direcţii de cercetare prezintă ideile principale de continuare a cercetării, printre care şi definirea unei metode care ajută la identificarea

scraper site-urilor, adică a site-urilor care copiază întreg conţinutul altor site-uri web cu scopul de a manipula motoarele de căutare şi de a atrage reclame.

Capitolul II - Interpretarea logurilor unei platforme de e-learning folosind Analiza Conceptuală Formală prezintă în cele patru secţiuni principale cum se poate observa comportamentul utilizatorilor în navigarea lor pe Internet cu ajutorul log-urilor unor servere web şi cum poate el determina performanţa acestora.

Secţiunea I - Analiza Conceptuală Formală prezintă aspectele teoretice din Analiza Conceptuală Formală (FCA), precum şi două ramuri ale acesteia: Analiza Conceptuală Formală Triadică (3FCA) şi Analiza Conceptuală Formală Temporală (TCA). De asemenea, tot aici sunt prezentate principalele instrumente utilizate de comunitatea cercetătorilor FCA.

Secţiunea II - Analiza comportamentului utilizatorilor web folosind Analiza Conceptuala Formala prezintă o nouă modalitate în care aceasta poate fi aplicată în *Web Usage Mining* pentru a determina dinamica utilizatorilor în portalul de *e-learning* studiat.

Secţiunea III - Vizualizarea datelor triadice prezintă CIRCOS, un instrument de vizualizare grafică a datelor obţinute folosind 3FCA. Scopul folosirii acestuia a fost de a evidenţia legăturile importante între conceptele triadice.

Secţiunea IV - Concluzii şi direcţii de cercetare prezintă ideile principale de continuare a cercetării, printre care şi construirea unui sistem adaptativ care se mulează pe descoperirea de noi informaţii în seturi de date folosind 3FCA şi TCA. Prin intermediul 3FCA se determina grupurile de utilizatori care tind să urmeze acelaşi tipar în vizitarea site-ului web, iar prin TCA se determină efectiv lanţurile de pagini care fac parte din traiectoriile cele mai des utilizate de către vizitatorii site-ului.

Lucrarea se încheie cu concluzii, o anexă ce cuprinde lista figurilor introduse în lucrare precum şi cu prezentarea bibliografiei utilizate pe parcursul redactării lucrării.

Cercetări anterioare în domeniu

Dată fiind actualitatea temei de cercetare propusă în această lucrare, literatura de specialitate prezintă o serie de publicaţii sau aplicaţii care o abordează.

Totodată, dinamica subiectului necesită o abordare continuă şi de substanţă.

Soluţiile adoptate în cercetările anterioare referitoare la subtemele propuse în această lucrare sunt:

- migrarea transparentă a unui site web:
  - existenţa unor *plugin*-uri implementate pentru majoritatea CMS-urilor care se bazează pe date statistice stocate de-a lungul timpului cu ajutorul cărora s-a stabilit un comportament al utilizatorilor, şi în funcţie de acesta se determină tiparul pe care e posibil să vrea să îl urmeze noul vizitator;
  - trimiterea unui utilizator nou către o pagină care dă eroare 404 sau la o pagină care are implementat un modul de căutare în interiorul site-ului web ([49], [17], [45], [20]);
- *spam linking* şi *spamdexing*:
  - dezvoltarea de tehnici anti *spam* ([38],[47]);

  - utilizarea tehnicilor de clasificare supervizată si nesupervizată în detectarea paginilor care sunt *spam*-uri
  - utilizarea metodelor învăţării automate ([47]);

- *Web Usage Mining*
  - folosind analiza statistică a datelor ([28]), graf-uri, instrumente care fac analiza logurilor web (i.e., Google Analytics, [16]) , business intelligence ([37], [39])

Rezultate şi contribuţii proprii

Dintre elementele originale propuse de autor, menţionăm următoarele:

- construirea *layer*-ului prin intermediul căruia se poate face migrarea transparentă de la un CMS la altul ([9]);
- alegerea unor măsuri de similaritate adecvate pentru algoritmul de regăsire a perechilor de URL-uri care prezintă conţinuturi similare;
- găsirea unei legături între rata de respingere a unei pagini dintr-un site web şi similaritatea acesteia cu *referrerii* care indică spre ea ([24]);
- aplicarea pentru prima dată a 3FCA în explorarea log-urilor web;
- utilizarea întregii suite FCA (3FCA şi TCA) pentru explorarea log-urilor web ([12]);
- utilizarea Circos, instrumentul de vizualizare grafică a datelor obţinute în urma folosirii 3FCA;
- determinarea dinamicii utilizatorilor folosind FCA şi 3FCA ([14]);

Capitolul 1

Impactul similarităţii web asupra traficului

1.1 Migrarea transparentă a unui site web între două sisteme de management de conţinut

Sistemele de management a conţinuturilor au început să fie din ce în ce mai utilizate în ultimul timp. Acest lucru reiese din top-ul realizat de site-ul alexa.com, în care 22,5% dintr-un milion de site-uri web sunt construite folosind un astfel de CMS, iar peste 50% dintre cele care folosesc un CMS utilizează Wordpress [33].

Avantajele utilizării unui astfel de sistem sunt: întreţinerea facilă a paginii web, independenţa conţinutului de prezentare, migrarea uşoară de la un design la altul, update-uri periodice de securitate, control total asupra optimizării motoarelor de căutare, administrare web, funcţionalităţi suplimentare oferite de terţi.

Argumentele aduse mai sus reprezintă o motivaţie clară a faptului că CMS-urile sunt din ce în ce mai folosite şi totodată aduc un plus migrării conţinutului site-urilor vechi la un astfel de sistem, în special în cazul în care CMS-ul curent nu mai poate îndeplini scopurile site-ului web, nu suportă noile obiective sau nu poate implementa toate funcţionalităţile, respectiv caracteristicile necesare îndeplinirii acestora.

1.1.1 Provocările procesului de migrare

Migrarea de la un CMS la altul poate fi făcută fie manual, fie automat. În cazul migrării de la un site static, procesul de migrare implică aproape în totalitate operaţii manuale. Pe de altă parte, migrarea de la un CMS la altul va implica operaţii automate.

Operaţiile care se fac în timpul procesului de migrare implică migrarea conţinuturilor paginilor web, dar mai mult decât atât cea mai dificilă parte a procesului de migrare este maparea vechiului conţinut celui nou corespunzător. De asemenea este importantă şi corelarea structurii organizatorice a vechiului site cu cea nouă dată de noul CMS.

Acest lucru poate avea o rezolvare graduală, urmărind trei paşi: împărţirea conţinutului în categorii, estimarea timpului necesar migrării, reevaluarea bazată pe ghidarea migrării.

Primul pas al procesului de migrare este clasificarea diverselor conţinuturi, în funcţie de rezultatul analizei acestuia şi de tipurile obţinute (prin tipuri se înţelege totalitatea categoriilor în care a fost subdivizat conţinutul). Astfel se nasc două tipuri de reguli

care trebuie identificate: acele reguli care se referă concret la conţinutul ales a fi necesar şi după migrare, precum şi regulile care se referă la ceea ce va putea şi la ceea ce nu va putea fi migrat automat. Aceste reguli care se definesc sunt utile în stabilirea priorităţilor şi în dozarea efortului care trebuie depus. În acest moment se poate decide ceea ce se poate migra automat şi ce nu. În general se doreşte să se automatizeze cât mai mult din ceea ce trebuie migrat.

Al doilea pas se referă la ghidarea şi estimarea timpului necesar migrării, adică compararea timpului necesar migrării automate, respectiv migrării manuale.

Ultimul pas presupune evaluarea procesului de migrare automată. Cea mai mare problemă care se pune atunci când se doreşte migrarea automată, este structura şi regularitatea conţinutului. Pe de altă parte, migrarea manuală spre un alt CMS presupune vizualizarea, editarea şi mutarea manuală a conţinutului, ceea ce probabil se va solda cu o risipire a resurselor de timp.

Procesul de migrare nu implică numai transformarea conţinutului vechi în conţinut nou, ci migrarea, fie ea automată sau manuală, ar trebui să ia în considerare comportamentul transparent al site-ului web, atât din punctul de vedere al vizitatorului sau al unui crawler al unui motor de căutare cât şi al unui browser web. O problemă comună care este indusă de o astfel de migrare este dată de expunerea conţinutului unui site web la un URL nou şi diferit faţă de cel vechi. După migrare, acest lucru se va reflecta în creşterea numărului de vizitatori care vor fi induşi în eroare de către link-urile apărute în *SERP (Search Engine Results Page)* sau de către terţi *referreri*. Pe lângă faptul că vizitatorilor nu li se afişează conţinutul dorit, site-ul ce tocmai a fost migrat ar putea pierde diversele beneficii câstigate în timp, pe care le avea datorită link-urilor postate pe diverse reţele de socializare sau a *back link*-urilor.

Astfel, a fost propus în [9] un *layer middleware* care să redirecţioneze transparent *request*-urile către URL-uri din vechiul site la URL-uri care aparţin noului site (atunci când maparea nu se poate face automat), pe baza, nu a conţinutului comun, ci pe baza semanticii conţinutului descris de cele două URL-uri şi pe baza semanticii unei informaţii adiţionale, cum ar fi query-ul dat de motoarele de căutare sau conţinutul de la URL-ul referrer-ului. Câteva dintre avantajele mecanismului propus, ar fi: reducerea numărului de *dead link*-uri venite dinspre referreri *3rd party*, asistarea motoarelor de căutare în direcţinarea corectă a utilizatorilor, precum şi conservarea *pageranking*-ului şi a rezultatelor din SERP. Spre deosebire de soluţiile propuse în cercetările anterioare ([49], [17], [45]), sistemul propus în [9] are avantajul de a fi implementat înainte de lansarea noului site, nedepinzând de date statistice strânse de-a lungul unei perioade de timp.

1.1.2 Tipuri de migrare a conţinutului

În continuare, vom propune un mecanism de formalizare a notaţiilor, adaptat CMS-urilor implicate în procesul de migrare (phpWebSite, Wordpress).

Conţinutul site-ului web este prezentat în moduri diferite în vechiul, respectiv în noul site. Astfel, apar două tipuri de conţinut ce se poate regăsi în structura site-ului: conţinut static (adică un fişier ce rezidă pe un server web, de exemplu o resursă .pdf sau .jpg) sau conţinut dinamic, preluat dintr-o bază de date.

Migrarea conţinutului static în conţinut static.

Chiar dacă URL-ul se schimbă numele fişierului rămâne la fel:

vechiul URL: `http://vechiulsite/vecheaadresa/numefisier`
noul URL: `http://vechiulsite/nouaadresa/numefisier`

Putem spune că similaritatea perfectă se obţine atunci când similaritatea dintre `URLvechi` şi `URLnou` este 1.

Migrarea conţinutul dinamic în conţinut dinamic.

Un astfel de conţinut se poate regăsi atât la un URl vechi, cât şi la unul nou, şi cuprinde două părţi: template-ul paginii web şi conţinutul stocat în baza de date.

Ceea ce face obiectul migrării este conţinutul absolut al paginii web, deoarece în calculul similarităţii dintre conţinuturile aflate în paginile de la url-ul vechi, respectiv nou, trebuie evitate informaţiile care ţin de template-ul site-ului şi care nu au nici o legătură cu conţinutul prezentat într-o anumită pagină.

Exemple:

- În Wordpress conţinutul absolut se poate extrage direct din baza de date, sau direct de la URL-ul asociat. Acesta poate fi diferenţiat faţă de template-ul paginii, datorită poziţionării într-un div care este unic determinat de id-ul `content`.

  Se vor nota în continuare paginile din vechiul site cu OP şi cele din noul site cu NP. Conţinutului absolut din site-urile construite în Wordpress sunt reprezentate de: pagini(atemporale), $NP_{\neg T}$, postări(temporale), NP_T, categorii, sau reuniune de postări şi submulţime de postări dintr-o categorie anume.

- În phpWebSite conţinutul absolut se regăseşte în: pagini cu o singură secţiune, OP_{SS} (migrate într-o pagină $NP_{\neg T}$) sau pagini cu mai multe secţiuni $OP_{MS} = \bigcup content(section_i)$. O astfel de pagină este migrată fie într-o singură pagină $NP_{\neg T}$, dacă secţiunile din paginea de la vechiul URL sunt atemporale, sau fiecare secţiune devine o postare NP_T ce va aparţine unei anumite categorii. Migrarea postărilor temporale în Word-press au ridicat anumite probleme, dat fiind faptul că în phpWebSite nu se reţin informaţii temporale legate de data creării sau modificării secţiunii din pagină.

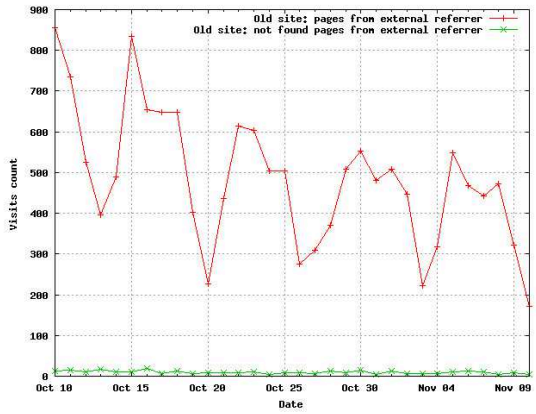
  S-au întâlnit dificultăţi la migrarea paginilor în postări temporale, datorită faptului că în phpWebSite nu sunt memorate date temporale.

Atunci când am comparat similaritatea conţinuturilor dintr-o pagină cu secţiuni multiple prezentată în vechiul site a fost nevoie să se facă compararea cu conţinutul absolut aflat într-o submulţime a postărilor dintr-o anumită categorie. Această necesitate va fi prezentată într-o secţiune ulterioară.

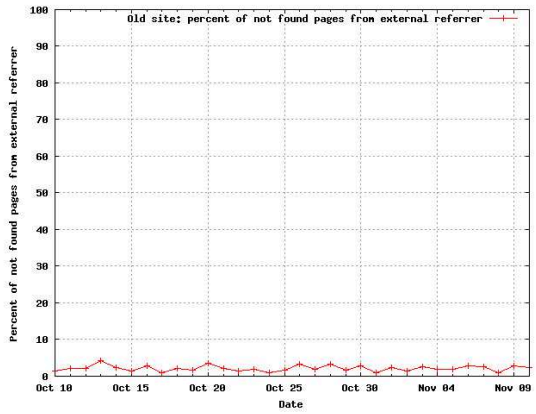
Migrarea conţinutul static în conţinut dinamic.

Această problemă apare atunci când e necesar să integrăm conţinuturile fişierelor HTML în structura generală a site-ului. Astfel, în acest caz, similaritatea se va calcula folosind conţinutul absolut din ambele site-uri.

Metoda propusă în cazul migrării unui site între două CMS-uri se bazează întotdeauna pe compararea conţinuturilor absolute, deoarece:



(a) Vechiul site: număr



(b) Vechiul site: procent

Figura 1.1: Pagini ce au generat eroare 404 şi au fost accesate de către un *referrer* extern

- este o metodă generală care poate fi luată în considerare atunci când se migrează şi de la un site static. Ea nu se bazează pe compararea efectivă a conţinuturilor din baza de date, ceea ce îi garantează caracterul general;
- alte metode nu acoperă cazurile enumerate anterior;
- este o metodă care nu necesită date de conectare la baza de date, datorită faptului că se bazează pe ceea ce este prezentat la un anumit URL. Acest lucru duce la ideea că nu este necesar să existe drepturi de administrator pentru că scopul este de a accesa cât mai uşor informaţia ce se regăseşte la un anumit URL;
- procesul prin care se regăsesc perechile corespunzătoare se face din perspectiva clientului.

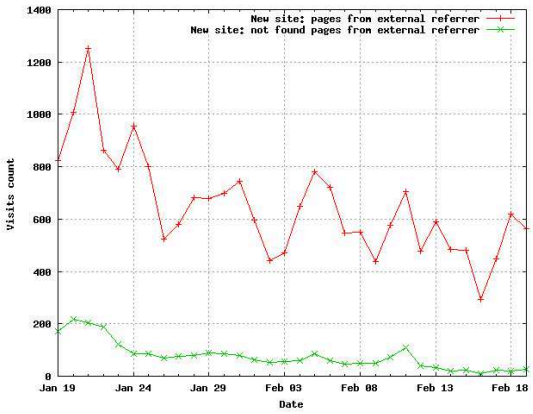
1.1.3 Implementare şi algoritm

Voi prezenta în continuare ideile principale ale algoritmului utilizat pentru a potrivi cât mai bine perechile de URL-uri din vechiul şi noul site. Pentru a găsi perechile potrivite, algoritmul utilizează o funcţie de similaritate a două şiruri de caractere şi anume similaritatea Cosinus. Algoritmul nu este dependent de funcţia de similaritate aleasă, însă utilizând această funcţie s-au obţinut cele mai bune rezultate în cel mai scurt timp [36].

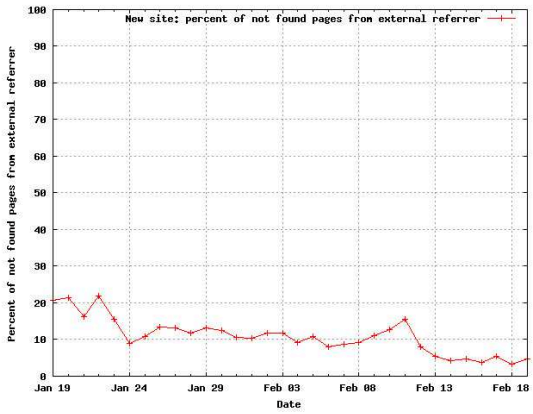
Procesarea informaţiei nu s-a făcut în timp real. Soluţia aleasă a fost aceea de a rula un program anterior accesării paginilor, astfel încât să se găsească cele mai bune potriviri între URL-uri. Am preferat acest lucru datorită faptului că atât în site-ul vechi, cât şi în cel nou, sunt mii de url-uri care trebuie luate în considerare, astfel, complexitatea algoritmului fiind egală cu cardinalul produsului cartezian al celor două mulţimi de URL-uri. Acest fapt ar duce la deservirea răspunsului către client cu întârziere. În plus, algoritmul propus este complet independent atât de platforma utilizată cât şi de limbajul de programare folosit în implementarea CMS-ului.

Prezentarea formală a algoritmului este următoarea:

```
Pentru fiecare URL din vechiul site la care se regăseşte conţinut static
  Identificarea URL-ului din noul site care indică spre acelaşi conţinut
```



(a) Noul site: număr



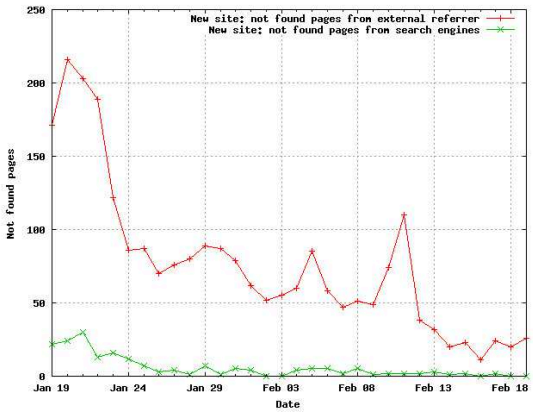
(b) Noul site: procent

Figura 1.2: Pagini ce au generat eroare 404 şi au fost accesate de către un *referrer* extern

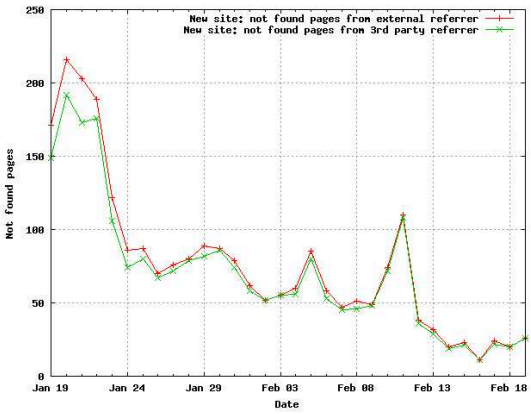
```
    acest conţinut se bazează pe numele fişierului
  Dacă URL-ul din noul site a fost identificat
    elimină vechiul URl din lista de URL-uri care trebuie să fie procesate
  SfDacă
SfPentru
Pentru toate URL-urile neprocesate
  Dacă la acest URL se regăseşte conţinut static
    conţinutul_absolut = conţinutul efectiv
  altfel
    conţinutul_absolut = conţinutul(vechiulURL) - conţinut (template-ul
        paginii)
  SfDacă
  Identifică URL-urile din noul site astfel încât conţinutul absolut
    are cea mai bună similaritate cu conţinutul absolut al paginii
    de la vechiul URl
  Perechea se inserează în baza de date, împreună cu:
    data, similaritatea obţinută la rularea algoritmului şi cea mai bună
    similaritate obţinută pe parcursul tuturor rulărilor
SfPentru
```

Conform algoritmului de mai sus s-au salvat în baza de date mai multe date referitoare la similaritatea perechilor de URL-uri, printre care şi data şi similaritatea obţinută la ultima rulare, precum şi cea mai bună similaritate obţinută de-a lungul timpului. Motivarea acestei alegeri vine de la faptul că postările dintr-o categorie evoluează în timp datorită noilor anunţuri ce pot să intervină şi astfel similaritatea obţinută la diferite rulări descreşte. Deci, vom avea similaritate maximă doar dacă comparăm conţinutul dintr-o anumită pagină cu secţiuni multiple (din vechiul site), cu o submulţime formată din mai multe anunţuri mai vechi (acestea vor corespunde cu cele mai noi anunţuri din noul site).

În acest sens prezint un exemplu: din categoria **Anunţuri Recente** din noul site, un vizitator este foarte posibil să îşi dorească să vizualizeze cele mai noi articole postate. Astfel, recdirecţionarea se va face spre acestea, şi nu spre un anunţ care se potriveşte



(a) Adaptarea motoarelor de căutare la noua structură a site-ului



(b) Majoritatea erorilor 404 sunt generate de către utilizatorii care vin de la referreri *3rd party*

Figura 1.3: Adaptarea rapidă a motoarelor de căutare la noua structură a site-ului

perfect, din punct de vedere al similarităţii,atunci când utilizatorul vine de la un referrer care face trimitere spre pagina cu anunţuri recente din vechiul site.

De asemenea, pentru a se grăbi puţin procesul de regăsire a perechilor, se poate determina un prag experimental, astfel încât orice pereche de tipul `(URL vechi, URL nou)` ce va avea similaritatea mai mare decât pragul experimental considerat, să se excludă din lista URL-urilor ce intră în procesul de comparare.

1.1.4 Rezultatele obţinute şi evaluarea acestora

Pentru a demonstra avantajele metodei propuse vom prezenta rezultatele unor experimente, efectuate pe parcursul a 30 de zile, care evidenţiază din punct de vedere numeric comportamentul noului şi respectiv al vechiului site atunci când vizitatorul doreşte să vizualizeze o pagină şi o accesează prin intermediul unui referrer extern (o anumită pagină este accesată prin intermediul motoarelor de căutare sau a altor site-uri care conţin link-uri spre site-ul migrat, si nu este accesată direct prin URL sau bookmark).

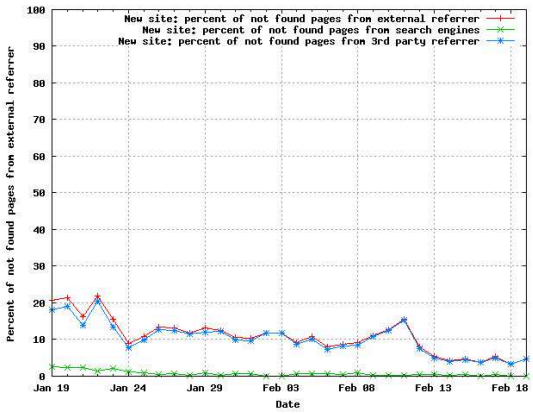
Astfel, au fost interesante rezultatele obţinute privind numărul paginilor negăsite, care au dat eroare 404 la cererea unei anumite resurse, precum şi procentul acestora relativ la numărul total de cereri HTTP care au fost efectuate.

Numărul paginilor care au generat eroare 404 înainte de migrare a fost relativ mic (1,9% Fig. 1.1(b)), acestea provenind în special de la referreri *3rd party*.

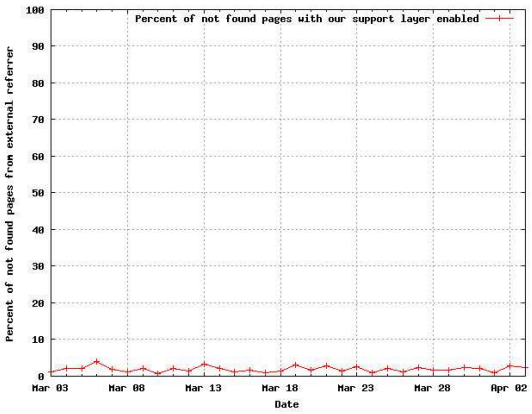
După migrare, numărul paginilor ce au generat eroare 404 şi au fost accesate prin intermediul unui *referrer* extern a crescut în mod natural la aproximativ 11.2% (Fig. 1.2(b)) pe parcursul celor 30 de zile considerate.

S-a observat că motoarele de căutare s-au adaptat repede la noua structură a site-ului, astfel, ele redirectând utilizatorii corect după o săptămână de la migrare (Fig. 1.3(a), Fig. 1.4(a)).

Procentajul paginilor ce generau eroarea 404, şi care aveau ca referreri motoarele de căutare, observate în prima săptămâna de după migrare era în jur de 1,9%, în timp ce după patru săptămâni, acest procent a scăzut la 0,23%. Acelaşi lucru s-a întamplat şi în cazul paginilor care aveau referreri *3rd party*, procentul scăzând de la 14.7% la 4.25% în patru săptămâni.



(a) Pagini ce generează eroarea 404 şi pagini ce generează eroarea 404 şi vin de la motoarele de căutare sau de la *referreri 3rd party*



(b) Pagini ce generează eroarea 404 având un *referrer* extern, odată ce a fost activat *layer*-ul suport propus

Figura 1.4: Adaptarea rapidă a motoarelor de căutare la noua structură a site-ului

Figura 1.4(b) prezintă rezultatele obţinute odată cu punerea în funcţiune a *layer*-ului suport propus în lucrarea de faţa, observându-se pentru site-ul ce tocmai a fost migrat, o îmbunătăţire a procentului de pagini ce generau eroarea 404 la 1.6%.

În algoritm s-a observat existenţa perechilor de URL-uri care aveau o similaritate de peste 70% în peste 93% din cazuri.

În concluzie, rezultatele experimentale prezentate mai sus au desăvârşit motivaţia conform căreia un astfel de *layer*, prin intermediul căruia se poate face migrarea transparentă de la un CMS la altul, este absolut necesar. Metoda propusă acoperă atât aspectele practice cât şi cele teoretice necesare mapării URL-urilor din vechiul site, celor corespunzătoare din noul site, pe baza similarităţii dintre conţinuturile prezentate la acele URL-uri. Totuşi, scopul introducerii acestui *layer* rămâne acela de a redirecţiona vizitatorul site-ului cu succes, indiferent de tipul referrer-ului de la care acesta vine.

1.2 Analizarea legăturii dintre similaritatea documentelor web şi a ratei de respingere generate de link-urile dintre aceste documente

Unul dintre principalele scopuri ale unui link este să ofere vizitatorului mai multe informaţii înrudite semantic cu informaţia din documentul considerat. Când vine vorba de link-uri externe de foarte multe ori în internet acestea sunt folosite abuziv, doar pentru a creşte *pagerank*-ul paginii sau al domeniului destinaţie şi nu pentru a ghida utilizatorul spre una sau mai multe pagini care să îi ofere în continuare informaţia de care să fie efectiv interesat.

În majoritatea cazurilor asemenea link-uri abuzive sunt fie *sitewide* (cel mai uşor de detectat), dar pot să fie plasate automat (în cadrul conţinutului absolut al unei pagini) de diferite module sau add-uri integrate în CMS-ul site-ului sursă.

Astfel, se pune problema existenţei unei posibile legături între similaritatea conţinutului documentului sursă cu conţinutul documentul indicat printr-un link extern prezent în sursă şi rata de respingere generată de link-ul respectiv în cadrul site-ului sau al domeniului destinaţie.

Rata de respingere asociată unei pagini web sau unui site reprezintă procentajul utilizatorilor care au vizitat o pagină din site şi au preferat să părăsească site-ul în loc să continue cu vizitarea altor pagini prezentate în cadrul aceluiaşi site. În general acest comportament se poate declanşa in urma a două posibile scenarii.

Primul motiv este acela că utilizatorul a găsit prezentat la acea pagină exact informaţia pe care o căuta. De exemplu, cineva poate căuta definiţia expresiei *rată de respingere*. Această informaţie este prezentată foarte bine chiar pe prima pagină care apare în SERP. El este mulţumit cu definiţia pe care a găsit-o şi părăseşte site-ul fără să mai acceseze şi alte pagini din interiorul acelui site.

Totuşi, în majoritatea cazurilor lucrurile se întâmplă diferit, utilizatorul nefiind mulţumit cu informaţiile oferite pe o pagină particulară din SERP, astfel părăsind-o pentru a putea accesa următoarele pagini din SERP.

Este bine cunoscut faptul că o rată de respingere mare, de obicei cauzată de al doilea caz prezentat anterior, are o semnificaţie negativă, iar aceasta este de obicei asociată cu calitatea conţinutului prezentat.

Cu cât documentele prezente pe domeniul extern oferă un conţinut mai similar cu documentul sursă, vizitatorului i se garantează accesul la tot mai multă informaţie care este de mare interes pentru el. Odată cu creşterea numărului de documente accesate ce aparţin site-ului destinaţie rata de respingere a acestuia scade.

Exemple:

- **forum dedicat iubitorilor de animale**

  Un astfel de forum poate conţine în unul din *topic*-urile deschise un link către un site ce se adresează crescătorilor de câini. Un astfel de link generează o rată de respingere mică, deoarece informaţia din pagina de destinaţie oferă un plus de conţinut de calitate, de care utilizatorul să fie interesat.

- **site-ul admiterii la Universitatea Babeş-Bolyai**
  Acest site prezintă informaţii detaliate despre procesul de admitere. Link-urile externe din orice pagină a acestui site reprezintă legături spre mai multe informaţii legate de admitere. Aceste informaţii sunt prezentate mai în detaliu pe site-ul corespunzător fiecărei facultăţi, site ce este găzduit pe un domeniu diferit. Astfel de link-uri oferă informaţii legate din punct de vedere semantic care sunt mult mai detaliate decât cele oferite pe site-urile admiterii, astfel generându-se o rată de respingere mai mică pentru site-ul facultăţilor.

Contraexemple:

- **site pentru care majoritatea link-urilor externe sunt reclame**

  Un link abuziv (spamlink, adlink, reclamă) în foarte multe cazuri duce spre site-uri cu un conţinut total diferit faţă de site-ul sursă, iar prin urmarea unui astfel de link se generează o rată de respingere mai mare.

Pentru a analiză această posibilă legătură se vor folosi mai multe funcţii de similaritate, precum: similaritatea Cosinus, similaritatea Jaccard, similaritatea Jaro-Winkler şi similaritatea Sorensen.

1.2.1 Rezultate anterioare

Odată cu creşterea exponenţială a Internetului şi a posibilităţii navigării pe web, expunerea link-urilor abuzive a devenit un fenomen negativ ce afectează calitatea rezultatelor returnate de către motoarele de căutare.

Din ce în ce mai multe companii din industria optimizării motoarelor de căutare, precum şi diverşi cercetători şi-au reunit forţele cu scopul de a identifica potenţiale soluţii referitoare la această problemă şi de a limita efectul negativ al acestui fenomen.

Încă de când analiza link-urilor a fost folosită pentru optimizarea motoarelor de căutare, s-a încercat introducerea de diverse tehnici de spammare [47], obţinându-se astfel multiple efecte negative ce au dus la noi provocări în ariile de cercetare.

Un studiu făcut asupra principalelor tehnici de detectare a link-urilor abuzive pe internet [38] clasifică aceste tehnici în trei clase: metode bazate pe analiza conţinutului, metode bazate pe analiza link-urilor şi metode bazate pe analiza datelor netradiţionale (sesiuni HTTP, comportamentul utilizatorului pe web). Prin aceste tehnici se detectează până la 80% din paginile abuzive [4], iar recomandarea autorilor este să se folosească tehnicile enumerate mai sus împreună, pentru a creşte procentul de pagini abuzive detectate (este recomandat să se folosească împreună cel puţin metodele ce analizează link-urile şi metodele care analizează conţinutul [3]).

Studiile anterioare care au propus diverse tehnici de detectare a *spam*-urilor web indică o posibilă clasificare automată a acestor tehnici prin metode de clasificare supervizate sau nesupervizate. De asemenea s-a mai propus şi utilizarea de algoritmi probabilistici, algoritmi de detectare a coliziunilor sau matrici de confuzie pentru anumiţi clasificatori[31]. Aceste propuneri au ridicat o mare provocare printre cercetători şi anume ridicarea *spamdexing*-ului la nivel de problemă de învăţare automată [47].

1.2.2 Prezicerea ratei de respingere folosind similaritatea conţinutului dintre sursă şi destinatie

Similaritatea ideală

Pe baza celor spuse mai sus, paginile care au o similaritate mare din punctul de vedere al conţinutului prezentat au şanse mai mici să genereze o rată de respingere mică, şi reciproc, paginile care au o similaritate mai mică a conţinutului generează o rată de respingere mai mare.

Astfel, problema care se pune este dacă este posibil să existe o legătură între rata de respingere şi rezultatele obţinute prin folosirea mai multor funcţii de similaritate.

Metodă propusă în [24] pentru a găsi o funcţie de similaritate ideală, ar putea să estimeze rata de respingere generată de link-uri de pe terţe site-uri spre terţe site-uri, fără a avea acces la diferite drepturi de administrare a domeniilor destinaţie (cum ar fi tool-urile oferite de Google Analytics).

O astfel de metodă ar permite identificarea pentru un site sursă a link-urilor abuzive, *add*-uri, impropriu plasate, care generează o rată de respingere mare, lucru care poate fi util în penalizarea site-ului sursă în cadrul motoarelor de căutare.

Totodată, o astfel de funcţie similaritate ideală ar trebui să se comporte ca şi în graficul 1.5, adică se poate face o interpolare liniară la funcţia de gradul I $f(x) = 100-x$.

Corelarea perfectă dintre rata de respingere şi similaritatea a două documente legate între ele (aşa cum este prezentat în cazul ideal, în figura 1.5), nu este întotdeauna găsită, în special luând în considerare funcţiile uzuale de similaritate. Astfel, am luat

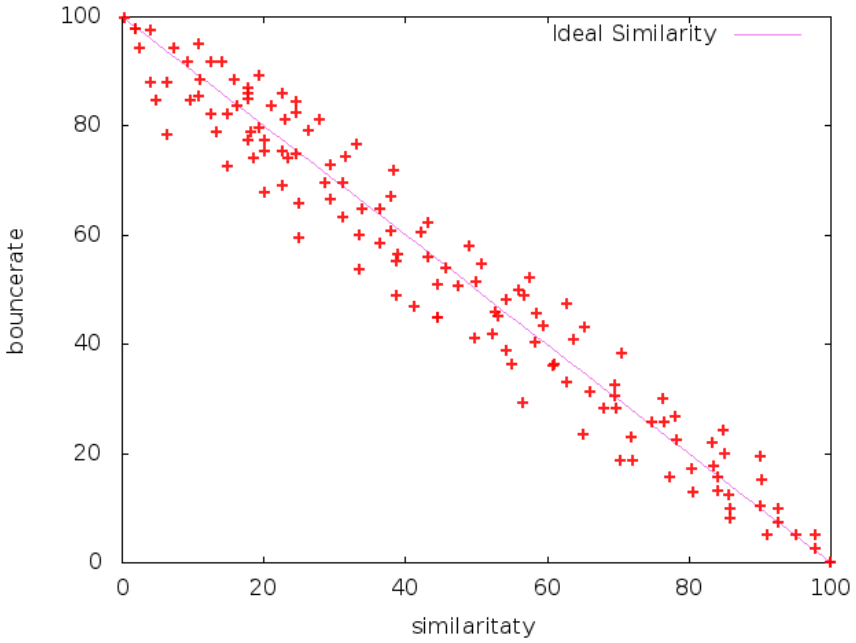


Figura 1.5: Similaritate ideală

în considerare mai multe astfel de funcţii de similaritate pentru a determina care dintre ele se potrivesc cel mai bine cu similaritatea ideală descrisă mai sus.

Intuitiv, se consideră că o funcţie de similaritate este mai bună decât o altă funcţie de similaritate, dacă perechile de tipul $(similaritate, rata\ de\ respingere)$, ce se obţin din link-urile ce leagă două documente, sunt mai aproape de diagonala reprezentată în cazul ideal în figura 1.5.

Pentru reprezentarea grafică a rezultatelor s-a considerat un reper cartezian xOy, unde similaritatea reprezintă abscisa punctului, iar rata de respingere reprezintă ordonata punctului reprezentat în sistemul considerat.

Luând în considerare dreapta paralelă cu a doua bisectoare ce trece prin punctele de coordonate (100,0) şi (0,100), o funcţie de similaritate este cea mai bună dacă suma tuturor distanţelor de la reprezentările grafice ale punctelor la dreapta mai sus menţionată este minimă, adică:

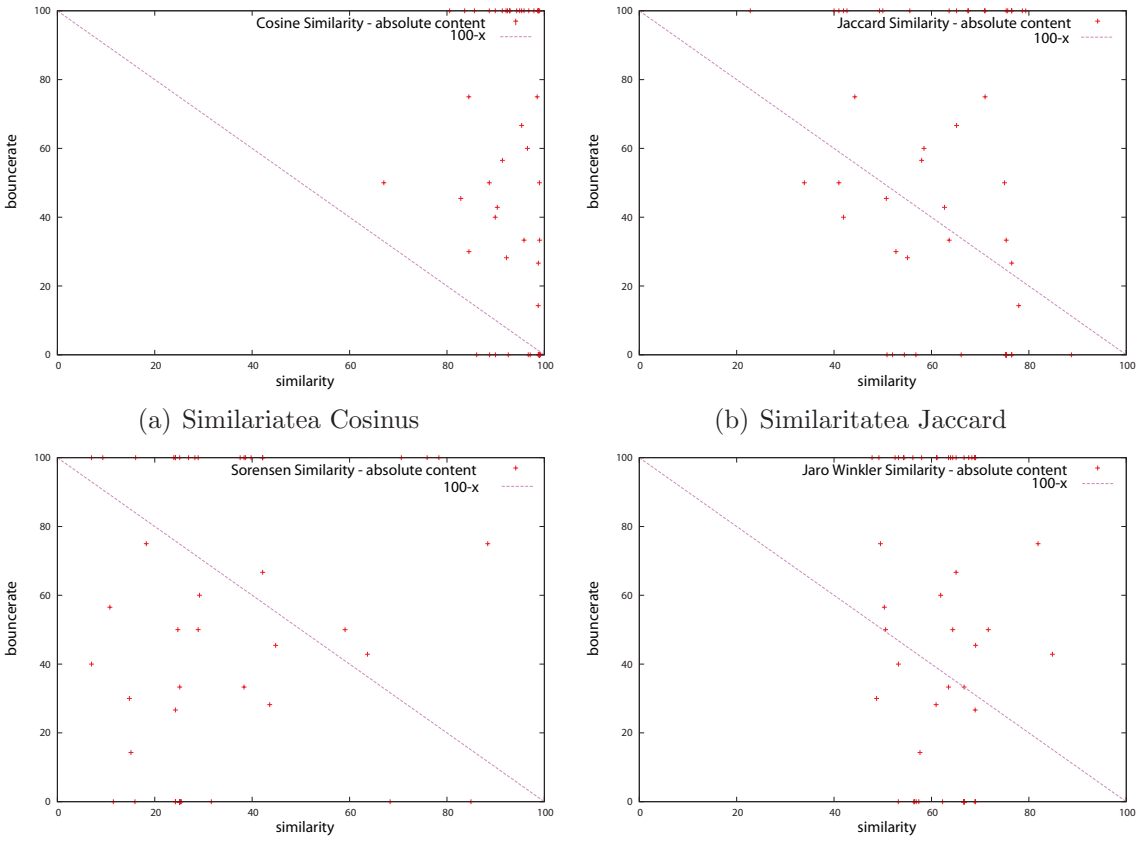
$$\sum \frac{|x_i + y_i - 100|}{\sqrt{2}}$$

este minimă, unde x_i şi y_i sunt coordonatele punctelor reprezentate grafic în sistemul cartezian de axe.

Rezultate experimentale

Pentru a evalua rezultatele experimentale obţinute prin metoda propusă, se vor lua în considerare: similaritatea Cosinus, similaritatea Jaccard, similaritatea Jaro-Winkler şi similaritatea Sorensen.

Pentru a obţine rezultate cât mai reale şi pentru a avea o acurateţe cât mai bună, am testat toate funcţiile de similaritate pe conţinutul absolut al paginilor web, ignorând conţinutul din şablonul paginii web, adică header-ul, footer-ul, respectiv meniul.



(a) Similariatea Cosinus

(b) Similaritatea Jaccard

(c) Similariatea Sorensen

(d) Similaritatea Jaro-Winkler

Figura 1.6: Similaritati

Conţinutul absolut al paginii web a putut fi complet determinat folosind o librărie Java, numită *boilerpipe*. Această librărie, oferită de Apache License 2.0, furnizează algoritmi care detectează şi elimină tot ceea ce ţine de template-ul site-ului web, păstrând doar conţinutul absolut al acestuia [1].

Pentru a demonstra ceea ce mi-am propus prin lucrarea de faţă, am luat în considerare site-ul Facultăţii de Matematică şi Informatică al Universităţii Babeş-Bolyai (http://www.cs.ubbcluj.ro) şi am generat tripletele (pagină accesată, referrer, rată de respingere), prin două metode:

- folosirea unui tool client-side oferit de Google (Google Analytics) ;
- folosirea unui tool server-side, integrat în template-ul site-ului, dezvoltat de autorii [24].

Reuşita experimentului se bazează pe faptul că am avut drepturi de administrare asupra site-ului mai sus menţionat, ceea ce a permis măsurarea corectă a ratei de respingere pentru fiecare link extern referrer al paginilor din acest site.

După cum se observă din figurile prezentate, cea mai bună funcţie de similaritate care oferă rezultatele dorite, este similaritatea Jaccard. Matematic, faptul că similaritatea Jaccard este cea mai bună funcţie de similaritate rezultă din calcularea sumei distanţelor de la fiecare punct ce e reprezentat grafic la dreapta de ecuaţie y=x-100.

Funcţie de similaritate	Metodă	Valoarea sumei
Cosinus	conţinut absolut	1804.476000579978
Jaccard	**conţinut absolut**	**1414.03085464388**
Sorensen	conţinut absolut	1769.3699189543242
Jaro-Winkler	conţinut absolut	1528.5359097516346

Tabela 1.1: Suma distanţelor de la toate punctele de pe grafic la dreapta de ecuaţie $y = x - 100$

1.3 Măsurarea şi vizualizarea nivelului de *scrapping* al unui site web

Odată cu dezvoltarea Internetului numărul site-urilor a crescut considerabil, şi astfel se poate explica uşor abundenţa informaţiilor care există pe web. Din păcate în ultimul timp atunci când se doreşte accesarea unei informaţii prin intermediul unui motor de căutare, atât user-ul cât şi motorul de căutare sunt puşi în situaţia unei probleme: prezenţa unor site-uri web în *SERP* care direcţionează greşit utilizatorii, sau îi direcţionează spre un site care preia complet informaţiile postate pe alt site. Acestea sunt numite *scraper site*s, iar de cele mai multe ori sunt considerate pagini web legitime.

În această categorie intră următoarele tipuri de site-uri:

- site-uri care publică întregul conţinut preluat de la un alt site, fără a-i adăuga nici o informaţie originală care să pună în valoare noua pagină;
- site-uri care copiază conţinuturile altor site-uri şi le postează într-o formă nouă, modificând automat anumite cuvinte din conţinut (un exemplu în acest sens ar fi folosirea sinonimelor);
- site-uri care reproduc informaţiile preluate din *feed*-urile *RSS (Rich Site Summary)* fără să adauge informaţii care sunt importante pentru vizitatorii site-ului
- site-uri care preiau fişiere multimedia (imagini, filme sau orice alt tip de media) şi le prezintă utilizatorilor fără a oferi un plus de informaţie vizitatorilor site-ului

Aşadar prin *scraper site* se înţelege un amalgam de conţinuturi ce au fost preluate din alte surse, de cele mai multe ori fără permisiune. Astfel de site-uri web sunt în general pline de reclame, iar scopul lor este să se interpună între utilizator şi site-ul care chiar dispune de informaţia pe care utilizatorul o caută. Astfel, *scrapper*-ul îşi atribuie de fapt rolul motorului de căutare, scopul său fiind acela de a creşte *pagerank*-ul *scraper*-ului în detrimentul site-ului original. Se doreşte atingerea acestui scop prin prezentarea frecventă a unui conţinut relevant şi unic preluat de la site-uri sursă care sunt bine punctate de motoarele de căutare.

Rezultatele mixte obţinute în *SERP* scad atât din performanţa motorului de căutare cât şi din mulţumirea utilizatorului referitoare la informaţia găsită. Dat fiind acest fapt, motoarele de căutare dezaprobă existenţa acestor tipuri de site-uri, tocmai din cauza faptului că acestea se interpun între user, motorul de căutare şi site-ul destinaţie.

Identificarea site-urilor de acest tip s-a incercat să se facă utilizând diverşi algoritmi, însă nu s-au obţinut rezultatele dorite. În consecinţă politica pe care unele motoare de căutare au anunţat-o la începutul anului 2014 a fost de a identifica *scraper*-ele prin intermediul utilizatorilor şi a *feedback*-ului pe care aceştia îl dau. Aceste site-uri odată catalogate nu sunt depunctate, ci clasificarea făcută de vizitatorii site-ului se utilizează în testarea algoritmilor prin intermediul cărora se detectează dacă un anumit site este *scraper* sau nu.

Totuşi această dezaprobare nu este întotdeauna una fermă, ci mai mult una declarativă, dat fiind faptul că unele motoare de căutare permit postarea de reclame pe care le furnizează pe *scraper site*-uri, astfel aducând un câstig în plus atât site-ului (deoarece apare în SERP), cât şi motorului de căutare (deoarece este un intermediar între cei care doresc să îşi facă reclamă prin *adlink*-uri şi cei care doresc să publice pe site astfel de reclame).

Pentru a îmbunătăţi calitatea rezultatelor oferite de motoarele de căutare, trebuie să se găsească o metodă de identificare a site-urilor care sunt *scraper*-e. Scopul identificării acestor site-uri este de a reduce, pe cât posibil, apariţia lor în *SERP*. Astfel, se doreşte în continuare studierea existenţei unei metode prin intermediul căreia aceste site-uri să poată fi identificate şi odată identificate acestea să fie penalizate de căutre motoarele de cătare astfel încat prezenţa lor în *SERP* să fie redusă cât mai mult posibil.

Datorită asistenţei aproape permanente de care un utilizator are nevoie atunci când îşi doreşte localizarea unor anumite informaţii pe web, calitatea acestor informaţii mai poate fi asigurată doar prin procesul de analizare a conţinuturilor prezentate pe web.

Pornind de la această idee, se va considera ca o direcţie de cercetare viitoare posibilitatea identificării unor astfel de pagini pe baza similarităţii dintre conţinutul aflat la pagina care face parte dintr-un *scraper site* şi pagina sursă de la care a fost preluat conţinutul. De obicei, pentru a nu încălca anumite reguli de confidenţialitate, *scraper site*-urile precizează pe fiecare pagină sursa de la care a preluat conţinutul, ceea ce face ca acestea să fie foarte uşor recunoscute "cu ochiul liber".

Astfel, în procesul de determinare a unui *scraper site* se vor lua în considerare mai multe site-uri, urmând etapele:

- se vor compara conţinuturile aflate în fiecare pagină din site-ul care se doreşte a fi testat cu conţinutul ce poate fi găsit urmând link-urile externe din pagina corespunzătoare,
- odată cu calcularea similarităţii dintre cele două conţinuturi, se va observa că pentru un *scraper site* se va obţine o similaritate mai mare în cazul în care link-ul extern se află în conţinutul absolut şi nu în template-ul paginii.

Pentru a studia fezabilitatea acestei idei am luat in considerare mai multe site-uri, care sunt actualizate zilnic. Cu ajutorul unui *crawler web* am indexat toate paginile site-urilor enumerate mai sus, după care, utilizând colecţia de date obţinută, am identificat în fiecare pagină conţinutul absolut prezentat precum şi sursa de la care a fost preluat conţinutul.

Aplicând etapele algoritmului descris mai sus, au fost generate tripletele *(link intern, link sursa, similaritate)*, pe baza cărora se va face identificarea *scraper site*-urilor.

Pentru a putea observa cât mai bine comportamentul unui astfel de site, pe baza tripletelor generate se vor reprezenta grafic punctele obţinute într-un sistem cartezian după următorul model:

- pe axa absciselor se va reprezenta similaritatea dintre cele două link-uri; în calculul acesteia s-a folosit măsura de similaritate a Cosinusului [1].

- pe axa ordonatelor se va reprezenta numărul de apariţii; prin acest număr de apariţii se înţelege numărul de perechi *(link intern, link sursă)* care au o similaritate ce se află în intervalul de încredere al similarităţii ce reprezintă abscisa punctului considerat. Printr-un interval de încredere al unui număr se va înţelege un interval de forma $[s - \epsilon, s + \epsilon]$, unde s este numărul căruia i s-a construit intervalul de încredere şi $\epsilon = 0.05$ este un prag determinat experimental.

Folosindu-ne de cele două coordonate ale unui punct de pe grafic, cu scopul de a evidenţia numărul de perechi pentru care s-a obţinut similaritate foarte mare între conţinutul regăsit la link-ul intern şi cel regăsit la link-ul sursă, pe reprezentarea grafică se regăsesc desenate cercuri pline cu centrul în punctul considerat şi cu raza direct proporţională cu numărul de apariţii.

În urma testelor efectuate se deduce faptul că pentru un *scraper site* graficul ce se obţine este mult *shiftat* în dreapta (Fig:1.7(a)). Aceste rezultate pot fi dovedite şi matematic, luând în considerare modul de construcţie a site-urilor de acest tip (similaritatea a două pagini (destinaţie, sursă) este direct proporţională cu numărul de apariţii).

Tocmai acesta este şi motivul pentru care s-a ales ca în reprezentarea grafică sa fie considerate şi cercurile de raze diferite. Dacă s-ar fi renunţat la această reprezentare, atunci multe dintre punctele aflate pe grafic ar fi fost suprapuse, datorită diferenţelor foarte mici dintre similarităţile corespunzătoare, iar efectul produs de *shiftare* la dreapta ar fi fost redus considerabil.

Totuşi, efectuând aceleaşi teste asupra unui site care citează pentru fiecare articol sursa şi care prezintă o informaţie modificată şi adaptată tipului de utilizatori căruia i se adresează, se observă că rezultatele obţinute sunt total diferite faţă de cele anterioare (Fig:1.7(b)). Graficul obţinut în cel de-al doilea caz indică un comportament normal şi chiar urmează curba lui Gauss, majoritatea punctelor de pe grafic aglomerându-se în centrul acestuia.

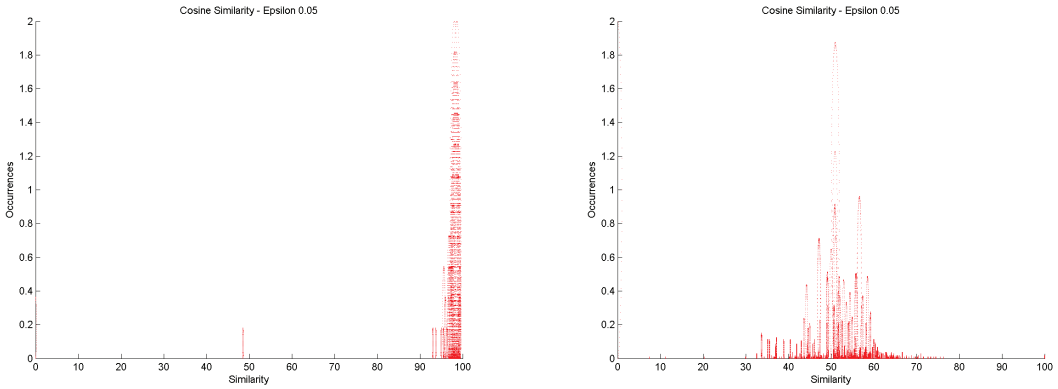
O analiză interesantă ar fi calcularea similarităţilor pentru mai multe astfel de site-uri şi sursele lor, observând astfel dacă rezultatul ce se va obţine se va putea generaliza pentru orice fel de site, iar în urma analizei rezultatului se va putea decide categoria în care acesta poate fi plasat.

1.4 Concluzii şi direcţii de cercetare

Prezentarea direcţiilor viitoare de cercetare începe cu prezentarea unor idei, care vin ca o continuare naturală a aspectelor prezentate în secţiunile anterioare.

O eventuală completare ce ar putea fi adusă *layer*-ului de migrare transparentă a conţinuturilor poate urmări mai multe aspecte:

- evaluarea diferitelor funcţii de similaritate, cu scopul de a descreşte timpul de regăsire a perechilor de URL-uri care prezintă conţinuturi similare;



(a) Scraper site

(b) Non-Scraper site

Figura 1.7: Similaritatea Cosinus - $\epsilon = 0.05$

- ponderarea unor anumite proprietăţi ale conţinutului prezentat la un anumit URL (cum ar fi: URL, *heading*-urile din pagină web: *h1,h2, ..., h6*, titlul paginii web, cuvintele cheie corespunzătoare care provin în urma accesării unui link din rezultatele afişate în SERP);

- redirecţionarea vizitatorului spre o pagină similară cu cea a *referrer*-ului.

În plu, este posibil să se obţină rezultate mai bune dacă se va considera analiza similarităţii conţinuturilor de la *referrer* cu conţinutul prezentat la orice link intern găsit în pagina accesată.

De asemenea, o alte direcţie de cercetare ar fi: utilizarea funcţiilor de similaritate care să interpreteze asemănarea dintre conţinuturi din punct de vedere semantic, ponderarea funcţiilor de similaritate sau alegerea unei funcţii de similaritate şi ponderarea unor proprietăţi specifice ale conţinutului.

Capitolul 2

Interpretarea logurilor unei platforme de e-learning folosind Analiza Conceptuală Formală

2.1 Analiza Conceptuală Formală

Analiza Conceptuală Formală (FCA) a fost dezvoltată constant pe parcursul ultimilor 30 de ani ([43]), pornind de la restructurarea teoriei laticilor. FCA cuprinde nu numai importante aspecte teoretice despre latici, despre proprietăţile lor şi despre structuri asemănătoare, ci şi algoritmi cu aplicaţii în *knowledge discovery* şi *knowledge representation*. De asemenea de-a lungul timpului au apărut şi diverse extinderi ale acestei teorii, precum: Fuzzy FCA, FCA Temporal, FCA Triadic, FCA relaţional.

În cazul datelor tridimensionale este necesară o abordare triadica.

Analiza Conceptuală Formală este un instrument matematic prin intermediul căruia se pot procesa conceptele.

Structura de bază este un *context formal*, din care se pot extrage şi procesa cunoştinţe folosind o conexiune Galois, numită operator de derivare.

Definiţie 2.1.1 *Un* ***context formal*** $\mathbb{K} := (G, M, I)$ *este format din două mulţimi* G *şi* M *şi o relaţie binară* I *între* G *şi* M*. Elementele din* G *se numesc* **obiecte** *şi elementele din* M *se numesc* **atribute***. Relaţia* I *se numeşte relaţie de incidenţă a contextului formal. Uneori se foloseşte notaţia:* gIm *în locul* $(g, m) \in I$*, pentru a exprima că obiectul* g *are atributul* m*.*

Definiţie 2.1.2 *Fie* $\mathbb{K} := (G, M, I)$ *un context formal. Pentru mulţimea de obiecte* $A \subseteq G$ *se defineşte:*

$$A' := \{m \in M \mid gIm \ \forall \ g \in A\}$$

mulţimea tuturor atributelor comune obiectelor din A*. Analog, pentru mulţimea atributelor* $B \subseteq M$ *se defineşte:*

$$B' := \{g \in G \mid gIm \ \forall \ m \in B\}$$

mulţimea tuturor obiectelor comune atributelor din B*.*

Conceptele sunt considerate unitatea de bază din care se pot desprinde cunoştinţe, iar ele sunt extrase din contextul formal utilizând operatorii de derivare.

Definiţie 2.1.3 *Un* **concept formal** *este definit ca o pereche* (A, B)*, unde* $A \subseteq G$*,* $B \subseteq M$ *şi* $A' = B, B' = A$*. Mulţimea* A *se numeşte* **extent** *şi conţine toate obiectele care au legătură cu conceptul, iar* B *se numeşte* **intent***, şi reprezintă mulţimea tuturor atributelor comune obiectelor din* A*.*

Mulţimea tuturor conceptelor dintr-un context dat (G, M, I) se notează $\mathfrak{B}(G, M, I)$. Se poate defini o ierarhie a conceptelor pe $\mathfrak{B}(G, M, I)$ folosind relaţia dintre subconcept-superconcept.

Definiţie 2.1.4 *Fie* (G, M, I) *un context. Relaţia* **subconcept-superconcept** *pe* $\mathfrak{B}(G, M, I)$ *este definită astfel:* $(A, B) \leq (C, D)$ *dacă şi numai dacă* $A \subseteq C$ *(sau echivalent,* $D \subseteq B$*), unde* $(A, B), (C, D)$ *sunt două concepte formale. Conceptul* (A, B) *se numeşte* subconcept *(sau* specializare*) pentru* (C, D)*, în timp ce* (C, D) *se numeşte* superconcept *(sau* generalizare*) pentru* (A, B)*.*

Relaţia subconcept-superconcept este o relaţie de ordine pe $\mathfrak{B}(G, M, I)$. În plus, $\mathfrak{B}(G, M, I)$ este o latice completă, numită **ierarhie de concepte**.

În astfel de diagrame de ordine, fiecare concept poate fi reprezentat într-un nod, iar relaţia de tip subconcept-superconcept este reprezentată sub forma unor muchii ce interconectează astfel de noduri.

Analiza Conceptuală Formală Triadică a fost introdusă în [29, 44].

Definiţie 2.1.5 *Un* context triadic *(sau:* tricontext*) este un cvadruplu* (K_1, K_2, K_3, Y)*, unde* K_1, K_2 *şi* K_3 *sunt mulţimi şi* $Y \subseteq K_1 \times K_2 \times K_3$ *este o relaţie ternară între ele. Elementele din* K_1, K_2, K_3 *sunt numite obiecte, atribute şi respectiv condiţii. Expresia* $(g, m, b) \in Y$ *se citeşte* obiectul g are atributul m luând în considerare condiţia b*.*

Următoarea definiţie arată cum contextele diadice pot fi obţinute dintr-un context triadic.

Definiţie 2.1.6 (Contexte derivate) *Pentru orice context triadic* (K_1, K_2, K_3, Y) *se pot obţine următoarele contexte diadice prin proiectare pe una din componente:*

$\mathbb{K}^{(1)} := (K_1, K_2 \times K_3, Y^{(1)})$ *cu* $gY^{(1)}(m, b) :\Leftrightarrow (g, m, b) \in Y$*,*

$\mathbb{K}^{(2)} := (K_2, K_1 \times K_3, Y^{(2)})$ *cu* $mY^{(2)}(g, b) :\Leftrightarrow (g, m, b) \in Y$*,*

$\mathbb{K}^{(3)} := (K_3, K_1 \times K_2, Y^{(3)})$ *cu* $bY^{(3)}(g, m) :\Leftrightarrow (g, m, b) \in Y$*.*

Pentru $\{i, j, k\} = \{1, 2, 3\}$ *şi* $A_k \subseteq K_k$*, se definesc* $\mathbb{K}^{(ij)}_{A_k} := (K_i, K_j, Y^{(ij)}_{A_k})$*, unde* $(a_i, a_j) \in Y^{(ij)}_{A_k}$ *dacă şi numai dacă* $(a_i, a_j, a_k) \in Y$ *pentru orice* $a_k \in A_k$*.*

Operatorii de derivare în cazul triadic se definesc folosind operatorii de derivare diadici pentru contextele diadice obţinute dupa proiectare.

Definiţie 2.1.7 (Operatorii de derivare - (i)**)** *Pentru* $\{i, j, k\} = \{1, 2, 3\}$ *cu* $j < k$ *şi pentru* $X \subseteq K_i$ *şi* $Z \subseteq K_j \times K_k$ *Operatorii de derivare-*(i) *se definesc prin:*

$X \mapsto X^{(i)} := \{(a_j, a_k) \in K_j \times K_k \mid (a_i, a_j, a_k) \in Y \text{ for all } a_i \in X\}$.

$Z \mapsto Z^{(i)} := \{a_i \in K_i \mid (a_i, a_j, a_k) \in Y \text{ for all } (a_j, a_k) \in Z\}$.

Bineînteles, aceşti operatori de derivare corespund cu operatorii de derivare din contextul diadic $\mathbb{K}^{(i)}, i \in \{1, 2, 3\}$.

Definiţie 2.1.8 (Operatorii de derivare - (i, j, X_k)**)** *Pentru* $\{i, j, k\} = \{1, 2, 3\}$ *şi* $X_i \subseteq K_i, X_j \subseteq K_j, X_k \subseteq K_k$*, Operatorii de derivare* (i, j, X_k) *se definesc prin:*

$X_i \mapsto X_i^{(i,j,X_k)} := \{a_j \in K_j \mid (a_i, a_j, a_k) \in Y \text{ for all } (a_i, a_k) \in X_i \times X_k\}$

$X_j \mapsto X_j^{(i,j,X_k)} := \{a_i \in K_i \mid (a_i, a_j, a_k) \in Y \text{ for all } (a_j, a_k) \in X_i \times X_k\}$.

Operatorii de derivare - (i, j, X_k) corespund operatorilor de derivare din contextul diadic: $(K_i, K_j, Y_{X_k}^{(ij)})$.

Similar noţiunii de cocept formal din contextul diadic, se pot introduce conceptele triadice.

Definiţie 2.1.9 *Un* concept triadic *(sau:* ***triconcept****) al* $\mathbb{K} := (K_1, K_2, K_3, Y)$ *este un triplet* (A_1, A_2, A_3) *cu* $A_i \subseteq K_i$ *pentru* $i \in \{1, 2, 3\}$ *şi* $A_i = (A_j \times A_k)^{(i)}$ *pentru orice* $\{i, j, k\} = \{1, 2, 3\}$ *cu* $j < k$*. Mulţimile* $A_1, A_2,$ *şi* A_3 *se numesc* ***extent, intent,*** *şi* ***modus*** *al conceptului triadic. Mulţimea* $\mathfrak{T}(\mathbb{K})$ *reprezintă mulţimea tuturor triconceptelor mulţimii* $\mathbb{K}$*.*

Propoziţie 2.1.1 *Triconceptele contextului triadic* (K_1, K_2, K_3, Y) *sunt tripletele maximale* $(A_1, A_2, A_3) \in \mathfrak{P}(K_1) \times \mathfrak{P}(K_2) \times \mathfrak{P}(K_3)$ *cu* $A_1 \times A_2 \times A_3 \subseteq Y$*, în raport cu relaţia de incluziune.*

2.2 Analiza comportamentului utilizatorilor web folosind Analiza Conceptuală Formală

Navigarea pe Internet are din ce în ce mai mult o dimensiune socială şi a devenit în ultimul timp un mecanism eficient de dobândire de cunoştinţe. Aşadar se impune necesitatea îmbunătăţirii atât a conţinutului prezentat utilizatorilor unui site web, cât şi a design-ului acestuia în special datorită faptului că au fost dezvoltate din ce în ce mai multe instrumente de analiză a comportamentului online al utilizatorilor.

Cercetările efectuate în ultimii ani relevă modul în care utilizatorii interacţionează pe Internet, impactul acestuia asupra vieţii de zi cu zi şi asupra culturii [41],[22],[7].

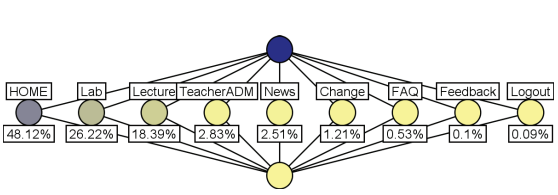
O atenţie specială trebuie acordată platformelor educaţionale. Acestea trebuie sa vină cu o structură clară şi cu un conţinut organizat.

Pentru a determina comportamentul utilizatorilor faţă de astfel de platforme online, sunt necesare colectarea de date adiţionale prin intermediul logurilor serverelor web.

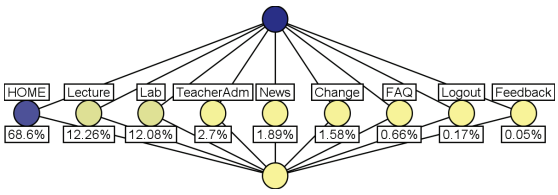
Metodele folosite până acum care implică extragerea de informaţii utile din aceste loguri web sunt metodele statistice sau tehnicile de *data mining*.

Instrumentele de analiză a logurilor web utilizează frecvent câteva metrici de bază. Însă acestea oferă o perspectivă reală în special site-urilor comerciale. Pentru site-urile educaţionale ele nu sunt eficiente, în special datorită faptului că procesul de învătare poate dura mai mult timp, iar în acest caz o vizită pe un site educaţional nu se supune euristicilor folosite de majoritatea instrumentelor de analiză [16].

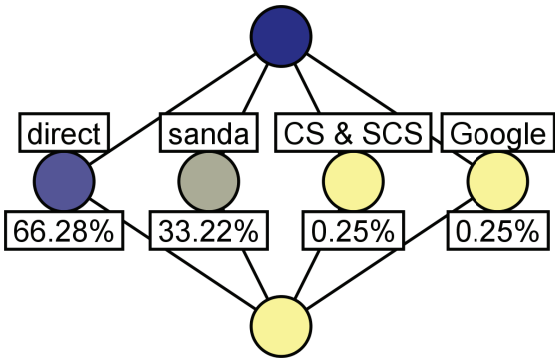
Analiza pe care ne-am propus să o facem foloseşte datele colectate de o platformă educaţională numită PULSE [15], în semestrul II al anului universitar 2012-2013. Pentru perioada considerată, au existat 40768 accesări pe platformă. Informaţiile colectate conţin informaţii despre: *URL*-ul complet al paginii accesate, *URL*-ul *referrer*-ului, *cookieID*, *login*, *timestamp*-ul accesării, adresa IP, *User Agent*, *Screen Resolution*. Informaţiile înregistrate sunt referitoare la două cursuri: Sisteme de Operare (SO1 -



(a) Clasele URL-urilor accesate



(b) 95.7% dintre *refereri* sunt din PULSE



(c) 4.3% dintre *referreri* sunt din afara PULSE

Figura 2.1: Scale nominale folosind ToscanaJ

curs obligatoriu) şi Proiectare Web şi Optimizare (WDO - curs opţional). Grupele de studenţi care au participat la cursul SO1 sunt notate "ar" şi "ri", iar grupele de studenţi care au participar la cursul WDO sunt notate "ei" şi "ie".

Am ales să nu folosim tehnicile clasice de data mining care ne asigură o perspectivă cantitativă, ci să avem o perspectivă calitativă prin utilizarea tehnicilor impuse de Analiza Conceptuală Formală, prin intermediul unui instrument numit ToscanaJ.

ToscanaJ [5] este un instrument utilizat îndeosebi pentru reprezentarea informaţiilor dobândite prin analiza datelor, şi oferă importante perspective asupra conexiunilor dintre datele analizate.

Pentru analizele efectuate am luat în considerare doar trei câmpuri din baza de date: *URL*-ul paginii accesate, *referrer*-ul şi *cookieID*-ul. Deoarece datele ce urmează a fi analizate sunt destul de consistente, este necesară pregătirea unei etape de preprocesare a acestora (există 751 de URL-uri, 471 de *referreri* şi 3472 *cookieID*-uri distincte).

Astfel, *URL*-urile accesate au fost clasificate în 9 clase disjuncte, care au fost vizualizate în ToscanaJ folosind o scală nominală (Figura 2.1(a)).

PULSE este o platformă care a fost construită în special pentru a fi folosită în timpul laboratoarelor, cu scopul de a consulta suportul teoretic oferit de că tre profesor, de a vizualiza problemele alocate sau de a vizualiza notele şi prezenţele. Fiecare vizitator PULSE se poate identifica în mod unic, deoarece este necesară trecerea printr-o etapă de autentificare înainte de a exista posibilitatea vizualizării oricărei informaţii enumerate anterior. Există două tipuri de utilizatori: studenţ şi profesor.

După autentificare, utilizatorul de tip student intră pe o pagină generală (din categoria *HOME*) care oferă informaţii generale despre: prezenţe, problemele de laborator alocate, suporturile teoretice aferente, note şi anunţuri curente.

Conform figurii 2.1(a) procentul celor mai multe accesări este atribuit clasei *HOME*. Acest lucru este de aşteptat dat fiind faptul că această clasă este un liant între toate celelalte clase. Următoarele clase frecvent vizitate sunt LAB şi LECTURE, care conţin materiale referitoare la laboratoare şi cursuri, exemple, lucrări, rezultate la lucrări,

precum şi soluţiile propuse pentru rezolvarea acestora.

Utilizatorului de tip profesor îi sunt atribuite drepturi administrative: atribuirea de probleme, acordarea de note, înregistrarea prezenţelor, postarea materialelor necesare, adăugarea de noutăţi. Toate aceste pagini intră în clasa TeacherADM.

Alte clase vizitate sunt NEWS (pagina pe care sunt publicate toate anunţurile), FAQ (pagina care conţine cele mai frecvente întrebări legate de platformă şi răspunsurile aferente), CHANGE (paginile cu materialele puse la dispoziţia studenţilor de la alte cursuri sau din anii anteriori) şi respectiv LOGOUT (pagina prin intermediul căreia se închide o sesiune pe PULSE).

Analog clasificării URL-urilor, s-a făcut o clasificare a *refererilor*, care reprezintă site-ul sau pagina de pe care au ajuns vizitatorii pe una din paginile din PULSE. *Referrerii* pot fi atât pagini din PULSE cât şi pagini din afara PULSE.

Folosind din nou ToscanaJ au fost vizualizate scalele nominale ale referrer-ilor din PULSE şi a celor din afara platformei (vezi figurile 2.1(b), 2.1(c)) .

Refererii interni au fost clasificaţi în aceleaşi clase ca şi URL-urile accesate. *Refererii* externi simbolizează fie accesările directe ale utilizatorilor (fie prin scrierea URL-ului, fie prin bookmark), fie accesările de pe reţele de socializare şi motoare de căutare (cum ar fi facebook, google), fie de pe site-ul facultăţii şi site-urile personale ale profesorilor.

2.3 Vizualizarea datelor triadice

Extensia Toscana2TRIAS permite selectarea datelor traidice, pornind de la o mulţime de scale preprocesate în ToscanaJ. Pentru a obţine un set de date triadice, am ales ca şi atribute clasele de *referreri*, ca şi condiţii clasele de URL-uri accesate, iar ca şi obiecte perechea (login, IP). Astfel am generat toate triconceptele folosind TRIAS.

Problema vizualizării datelor triadice nu a fost încă rezolvată satisfăcător, singurele opţiuni disponibile fiind trilaticile sau graf-urile. Din aceste motive am ales să folosim CIRCOS, un instrument nou de vizualizare a datelor, care a fost construit tocmai pentru a ajuta la investigarea tiparelor în date.

CIRCOS este un instrument prin intermediul căruia datele şi conexiunile dintre ele pot fi vizualizate într-un format circular [10]. Datele de intrare pentru CIRCOS au fost obţinute din tricontext utilizând operatorii de derivare. A fost necesară implementarea unui algoritm care transformă XML-ul rezultat din TRIAS în formatul valid de intrare acceptat de CIRCOS.

XML-ul rezultat din TRIAS conţine toate triconceptele care pot fi derivate din tricontextul definit. Fiecare triconcept este definit de către *extent*, *intent* şi *modus*. Formatul datelor de intrare acceptat de CIRCOS este un tabel bidimensional $R \times C$, care conţine valori numerice. Acesta a fost obţinut după cum urmează:

Pornind de la tricontextul (G, M, B, Y), am construit proiecţia diadică $\mathbb{K}_{32} := (G, (B, M), I)$, unde $(g, (b, m)) \in I \Leftrightarrow (g, m, b) \in Y$. Dupa aceea, pentru fiecare pereche (b, m) am evaluat conceptul corespunzător $\mu_{\mathbb{K}_{32}}$ şi am determinat cardinalitatea *extent*-ului $(b, m)'$.

Mulţimea care defineşte coloanele tabelului, notată C, este mulţimea obţinută prin proiectarea relaţiei de incidenţă Y pe M, $\mathrm{pr}_M(Y) := \{m \in M \mid \exists (g, b) \in G \times C.\ (g, m, b) \in Y\}$. Analog, se obţine mulţimea care defineşte liniile tabelului, notată R, şi se obţine prin proiectarea relaţiei de inciden tă Y pe mulţimea condiţiilor B.

Algoritmul implementat, construieşte un tabel care are ca indicatori de coloană şi de linie mulţimile construite anterior, şi calculează valorile numerice din tabel în modul în care urmează. Pentru fiecare pereche $(c, r) \in C \times R$, numărul de elemente din *extent* $(c, r)' \in \mathbb{K}_{32}$ este calculat direct din XML-ul rezultat din TRIAS. Cardinalul acestei mulţimi reprezintă valoarea numerică ce se va regăsi în tabel la intersecţia coloanei c cu linia r.

Exemplu de XML generat din TRIAS

```
<Triconcepts>
  <Triconcept id="1">
    <Extent>
      <Object>e1</Object>
      <Object>e2</Object>
    </Extent>
    <Intent>
      <Attribute>i1<Attribute>
      <Attribute>i2<Attribute>
    </Intent>
    <Modus>
      <Condition>m1</Condition>
    </Modus>
  </Triconcept>
  <Triconcept id="2">
    <Extent>
      <Object>e1</Object>
    </Extent>
    <Intent>
      <Attribute>i2</Attribute>
    </Intent>
    <Modus>
      <Condition>m2</Condition>
      <Condition>m3</Condition>
    </Modus>
  </Triconcept>
</Triconcepts>
```

Exemplu de date de intrare formatate pentru CIRCOS

```
-   i1  i2
m1  2   2
m2  0   1
m3  0   1
```

Figura 2.2 prezintă un exemplu de format circular generat de CIRCOS folosind o mulţime de triconcepte generate de TRIAS având ca şi obiecte:*cookieID*, ca şi atribute: clasele de *referreri* şi condiţii: clasele de URL-uri accesate. Deoarece clasele de *referreri* (R_CLASS) şi clasele de URL-uri accesate (AF_CLASS) au elemente în comun, mulţimile C şi R nu sunt disjuncte. În figura 2.2 segmentele sunt reprezentate circular, în sensul acelor de ceasornic, fiind ordonate descrescător, începând cu segmentul verde "HOME", segmentul mov "LECTURE", segmentul albastru "LAB" şi segmentele "NEWS", "CHANGE", "FAQ", "ADM", "FEEDBACK" and "LOGOUT". Reprezentarea circulară obţinută poate fi interpretată conform explicaţiilor din Figure 2.2.

Fiecare bandă corespunde unei perechi (R_CLASS, AF_CLASS). Culoarea benzii este dată de culoarea segmentului *referrer*-ului. De exemplu banda verde de la "HOME" la "LECTURE" corespunde tuturor paginilor accesate din clasa "LECTURE" de la *referreri* din clasa "HOME". Pentru o întelegere mai bună a conexiunilor, am reprezentat cu ajutorul unui graf orientat (Figura 2.3) conexiunile dintre clasele de *referreri* şi de URL-uri accesate. Nodurile graf-ului au culorile corespunzătoare benzilor din reprezentarea grafică rezultată din CIRCOS. Deoarece mulţimile claselor de *Referreri* si de URL-uri accesate nu sunt disjuncte, vor exista atât muchii orientate în ambele direcţii (înspre şi dinspre acelaşi nod), precum şi cicluri.

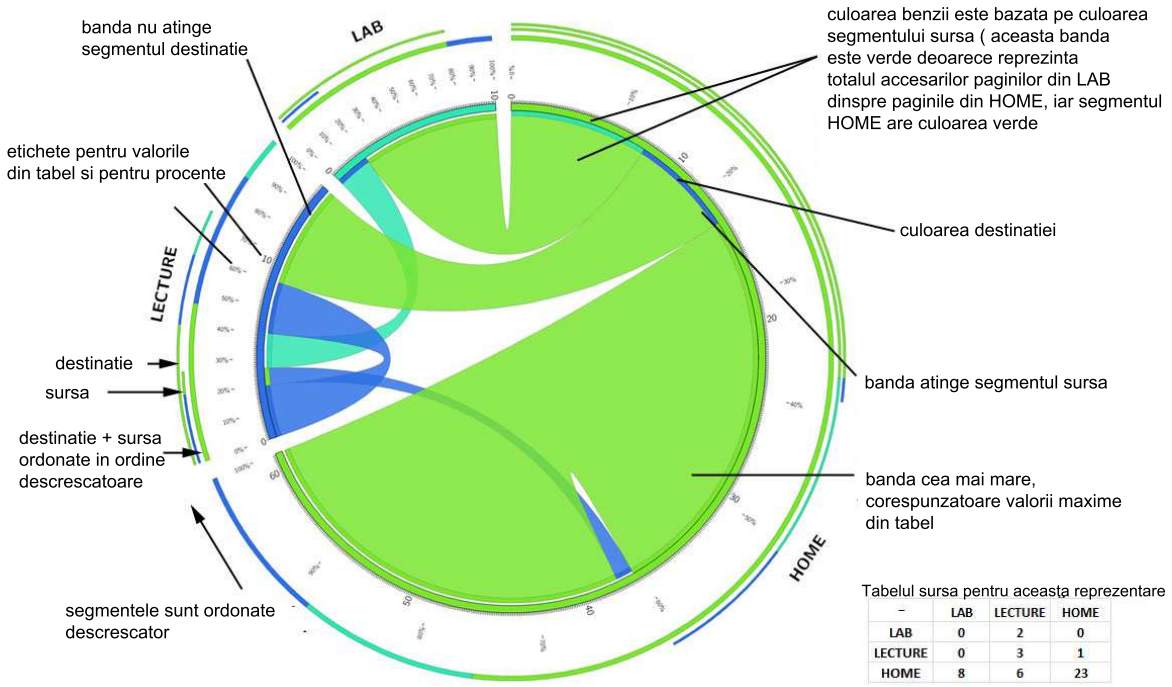


Figura 2.2: Objects Set: CookieId, Attribute Set: Referrer Class, Conditions Set: Access File Class

Luând în considerare datele logate de PULSE, am cautat diferite structuri triadice în acestea, pentru a putea fi interpretate folosind Toscana2TRIAS. Dintre toate testele efectuate, rezultate reprezentative pentru tiparuri în traiectoria studenţilor în navigarea lor prin site [12].

Testele au fost efectuate luând ca şi obiecte: *cookieID*, atribute: R_class şi condiţii AF_class. Am considerat doar *referreri*i care sunt din interiorul PULSE (Figura 2.4(a)).

Se observă că cele mai multe accesări sunt făcute din clasa HOME spre clasa HOME. Acest lucru se explică prin tranziţiile de la pagina de login la pagina principală, prin selectarea anului de studiu şi a cursului pentru care utilizatorul doreşte să vizualizeze informaţiile sau reîncarcarea paginii.

După ce sunt urmaţi aceşti paşi, cele mai multe vizite sunt de la HOME la paginile din clasele LAB, LECTURE sau NEWS. Aceste rezultate dovedesc scopul principal al platformei educaţionale PULSE şi anume deservirea de suport teoretic pentru laboratoare, cursuri şi postarea de anunţuri.

Accesarea paginilor din clasa LECTURE seminifică faptul că utilizatorii au vizitat atât slide-urile de la curs cât şi informaţiile despre lucrările care s-au dat pe parcursul

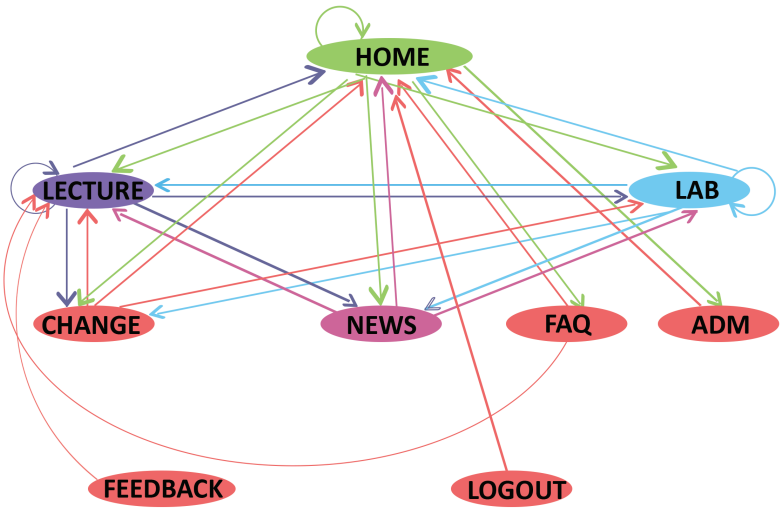
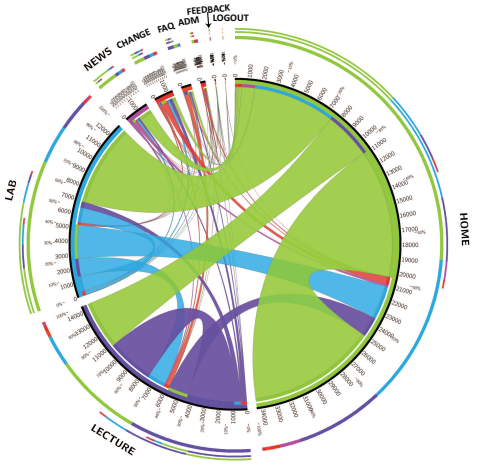


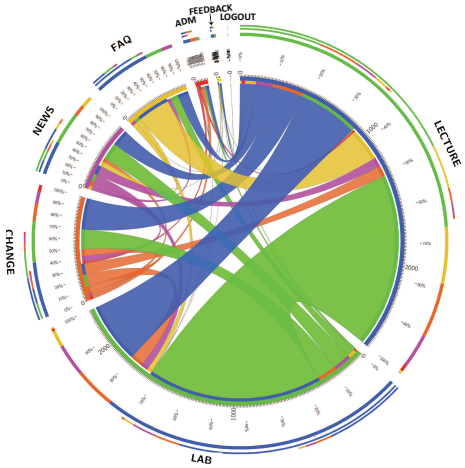
Figura 2.3: Vizualizarea cu ajutorului graf-ului orientat a conexiunilor dintre clasele de *referreri* şi URL-uri accesate

semestrului sau explicaţii detaliate despre rezolvarea lor. Aceste tiparuri în navigare în aceeaşi clasă sunt naturale datorită modului de construcţie al claselor (fiecare clasă conţine mai multe pagini). Mai mult decât atât aceste tiparuri se regăsesc în toate clasele existente.

Cu scopul de a observa şi alte pattern-uri am eliminat din datele de intrare pentru CIRCOS astfel de legături. De asemenea am eliminat toate interacţiunile dintre HOME şi alte clase (Figurile 2.4(a), 2.4(b)).



(a) Interpretarea trend-urilor triadice deduse din triconcepte, considerând R_CLASS, AF_CLASS şi cookieID



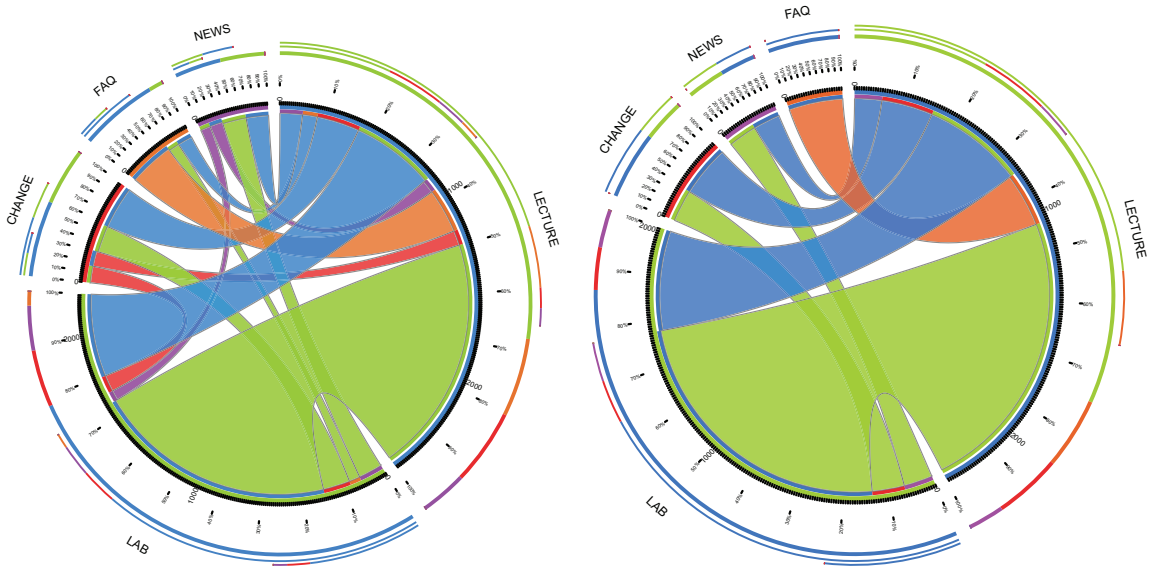
(b) O nouă interpretare a Figurii 2.4(a) după eliminarea interacţiunilor

Figura 2.4: Interpretarea trend-urilor triadice deduse din triconcepte

După eliminarea elementelor de mai sus, se pot observa şi alte tipare în navigarea

studenţilor prin platformă: acticitate intensă de la paginile cu laboratoare la paginile cu cursuri, sau reciproc de la cursuri la laboratoare, utilizarea facilităţii de CHANGE, care e cel mai adesea accesată de către studenţii din paginile din clasele LAB şi LECTURE, şi reciproc, vizitarea paginilor din clasele LAB şi LECTURE din paginile din clasa FAQ
.

Alte detalii sunt dificil de observat luând în considerare aceste reprezentări. Aşadar, pentru a vedea trend-urile cele mai importante între clasele principale din PULSE, am eliminat gradual valorile (numărul de vizite) mai mici decât 50 şi respectiv 100, considerând că nu sunt foarte reprezentative dat fiind faptul că valoarea maximă este mai mare decât 1000. Rezultatele se pot observa în Figura 2.5.



(a) Peste 50 de vizite

(b) Peste 100 de vizite

Figura 2.5: O noua interpretare a Figurii 2.4(b) fără a lua în considerare clasele cu prea puţine vizite

Principalele clase din PULSE sunt: LECTURE, LAB, NEWS, CHANGE şi FAQ. Figura 2.5(a) arată că sunt mai mult de 50 de vizite dinspre LECTURE către orice altă clasă. Acelaşi lucru se poate spune şi despre LAB. Pentru CHANGE şi NEWS, interacţiunile sunt bidirecţionale, dar doar cu clasele LAB şi LECTURE, î timp ce înspre FAQ vizitatorii pleacă atât de la LAB cât şi de la LECTURE, dar se întorc la FAQ doar de la LECTURE.

Pentru interacţiunile cu cel puţin 100 de vizite, direcţia de vizitare la la LAB sau LECTURE este unidirecţionala către CHANGE şi NEWS.

Numărul mare de accesări de la FAQ la LECTURE a fost surprinzător. Investigând motivul pentru care studentţii urmează acest tipar am ajuns la concluzia că pagini din cele două clase sunt consecutiv amplasate în meniul de navigare, prin urmare există posibilitatea ca utilizatorii care au ajuns pe pagina de FAQ de multe ori să fi ajuns din greşeala. Aşadar se impune o modificare a design-ului paginii în acest sens.

Pentru a continua abordarea triadică am luat în considerare R_CLASS ca şi obiecte, AF_CLASS ca atribute şi timestamp ca şi condiţii. Pentru a investiga comportamentul temporal al studenţilor pe parcursul unui semestru am analizat datele pe intervale de timp [12].

Primele intervale de timp considerate au fost egale cu aproximativ o treime din

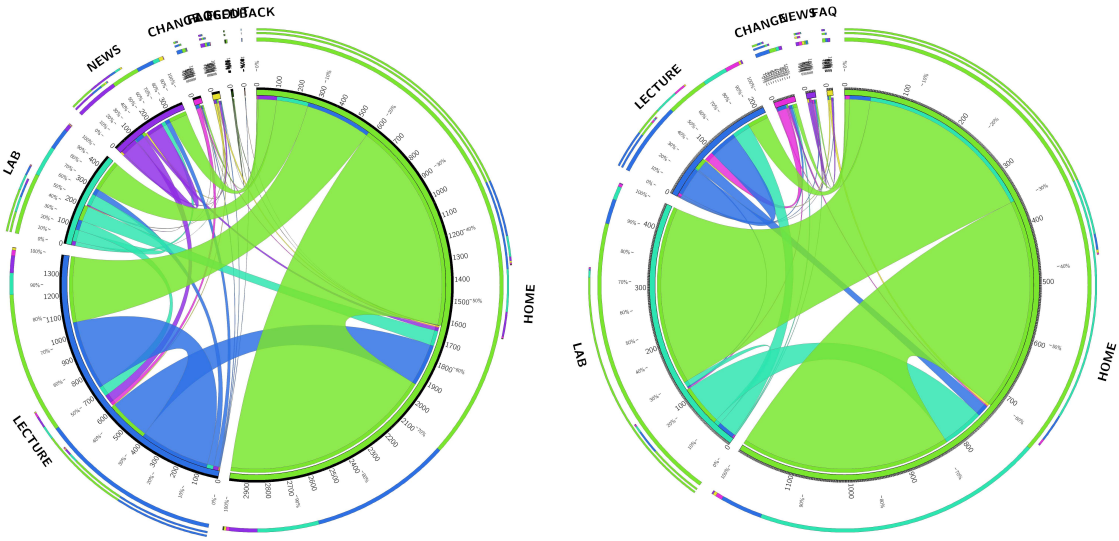
semestru, şi au fost denumite *început*, *mijloc* şi *sfârşit*. Astfel de intervale au conţinut un volum prea mare de date, şi astfel nu au generat trend-uri semnificative în datele considerate.

Aşadar am considerat intervalele de timp săptămânile din semestru. Lista completă a rezultatelor obţinute se regăseşte pe `http://www.cs.ubbcluj.ro/~fca/tests-2013`. În această analizăm observat trei tendinţe de comportament: relaxat, intens şi normal.

Comportamentul **relaxat** apare în principal în timpul vacanţei (de exemplu săptămâna 10). Tiparul urmat poate fi vizualizat în Figura 2.2 şi poate fi distins datorită numărului mic de URL-uri accesate (de obicei doar URL-urile din clasele HOME, LAB, LECTURE). Traiectoriile de navigare prin site observate pe parcursul acestei săptămâni sunt simple: vizitarea paginilor dinspre HOME spre LAB sau LECTURE; dinspre LAB spre LECTURE, dinspre LECTURE spre HOME. Pentru cursul opţional (WDO) rezultatele evidenţiază traiectorii de navigare mult mai relaxate datorită faptului că acest curs implică cercetare. Aşadar, materialul de curs nu este la fel de mult vizitat ca în cazul cursului obligatoriu (SO1). Acest tip de comportament apare şi dupa examenele finale (săptămânile 18 şi 20 pentru grupa "ar", 18 şi 19 pentru grupa "ri" şi săptămâna 14 pentru grupele "ei" şi "ie").

Comportamentul **intens** apare în timpul perioadei de examinare (săptămânile 17,19 şi 20 pentru grupele "ar" şi "ri" şi săptămânile 7,9 şi 13 pentru grupele "ei" şi "ie"). Traiectoriile utilizatorilor pot fi desprinse vizualizând Figura 2.6(a) care prezintă un număr ridicat de accesări. Aceste tiparuri pot apărea şi în saptămânile de dinaintea examinării.

Comportamentul **normal** apare pe parcursul semestrului, atunci când nu este perioadă de examinare sau vacanţă. Tiparul observat este evidenţiat în Figura 2.6(b), adică pagini din aproape toate clasele au fost vizitate.



(a) Săptămâna 17 - intens

(b) Săptămâna 5 - normal

Figura 2.6: Compararea comportamentului grupei "ar": intens versus normal

Deşi în Figura 2.6 cele două comportamente par similare, există diferenţe semnificative, şi anume faptul că în perioada cu comportament intens paginile din clasa LECTURE sunt cele mai vizitate, dat fiind faptul că studenţii se pregătesc pentru examene, iar în perioada cu comportament normal paginile din clasa LAB sunt cele

mai vizitate, dat fiind faptul că studenţii trebuie să îsi rezolve probleme de laborator pe care le-am primit. O altă diferenţă ar fi chiar numărul de accesări al paginilor din clasa HOME. În perioada cu comportament intens, clasa HOME pare a fi un liant între celelalte facilităţi oferite de PULSE.

Aşadar, informaţiile oferite de această interpretare, bazată pe triconcepte sunt calitative, spre deosebire de diverse instrumente statistice (cum ar fi histogramele) care oferă informaţii cantitative.

2.4 Concluzii şi direcţii de cercetare

Ca direcţie de cercetare viitoare ne propunem ajustarea parametrilor în fişierele de configurare ale CIRCOS: ajustarea ordinii de afişare a benzilor, ajustarea transparenţei benzilor în funcţie de distribuţia valorilor în datele de intrare, ascunderea sau ştergerea unor benzi care nu corespund unor condiţii date, scalarea valorilor din celule.

În plus, plănuim să alegem şi alte formate de reprezentare grafică în afară de cel circular.

O altă problemă pe care am vrea sa o abordăm în cercetările viitoare este numărul mare de concepte generate de Trias. Pentru a rezolva această problemă propunem construirea unui algoritm bazat pe probabilităţi condiţionale pentru a şterge conceptele cele mai puţin importante.

De asemenea, componenta temporală este foarte importantă în log-urile web. Ne propunem să folosim Analiza Conceptuală Formală Temporală pentru a studia comportamentul utilizatorilor, atât individual cât şi pe grupuri.

Listă de figuri

Bibliografie

[1] Apache Lucene,A high-performance, full-featured text search engine library, *http://lucene.apache.org/*

[2] Apache HTTP Server, *http://httpd.apache.org/docs/current/mod/ mod_alias.html*

[3] L. Becchetti, C. Castillo, D. Donato, *Link-Based Characterization and Detection of web Spam*, 2nd International Workshop on Adversarial Information Retrieval on the web, AIRweb, Seattle, USA, August 2006, pp. 1-8

[4] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, R. Baeza-Yates, *Link Analysis for web Spam Detection: Link-based and Content Based Techniques*, ACM Transactions on the web (Tweb), Volume 2, Issue 1, New York, USA, February 2008, pp. 1-41

[5] P. Becker, J. Hereth, G. Stumme, *ToscanaJ: An Open Source Tool for Qualitative Data Analysis, Advances in Formal Concept Analysis for Knowledge Discovery in Databases.*, page 1-2. Lyon, France, (July 2002).

[6] B. Berelson, *Content analysis in communication research*, New York, Free Press, 1952

[7] T. Berners-Lee , W. Hall, J. Hendler, N. Shadbolt, D.J. Weitzner, *Science 313*, 769 (2006).

[8] G. Beydoun, R. Kultchitsky, G. Manasseh, *Evolving semantic web with social navigation, Expert Systems with Applications*, 32(2007), pp. 265–276.

[9] D. Bufnea, **D. Haliţă**, *A server-side support layer for transparent web content migration*, Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT 2013 Cluj-Napoca, Studia Universitatis Babeş-Bolyai Informatica, 3/2013, pp. 78-89

[10] Circos, a circular visualization tool, www.circos.ca

[11] A. Dieberger, *Supporting social navigation on the world wide web.*, International Journal of HumanComputer Studies, 46(6), 805825, 1997.

[12] Sanda Dragoş, Diana Haliţă, Christian Săcărea, Diana Troancă, *An FCA grounded study of user dynamics through log exploration*, Studia Universitatis Babes-Bolyai Series Informatica, Volume LIX, no. 2, 2014, pp. 82-97

[13] S. Dragoş, C. Săcărea, *Analysing the Usage of Pulse Portal with Formal Concept Analysis*, Studia Universitatis Babes-Bolyai Series Informatica, LVII (2012), pp. 65–75.

[14] Sanda Dragoş, Diana Haliţă, Christian Săcărea, Diana Troancă *Applying Triadic FCA in Studying Web Usage Behaviors*, Knowledge Science, Engineering and Management, 7th International Conference, Volume 8793, 2014, pp. 73-80

[15] S. Dragoş, *PULSE Extended, in The Fourth International Conference on Internet and Web Applications and Services*, Venice/Mestre, Italy, May 2009, IEEE Computer Society, pp. 510–515.

[16] S. Dragoş, *Why Google Analytics can not be used for educational web content*, 2011 International Conference on Next Generation Web Services Practices.

[17] Dynamic404, Logical error-pages, *http://www.yireo.com/software/joomla-extensions/dynamic404*

[18] M. Eirinaki, M. Vazirgiannis, *Web mining for web personalization*, ACM Transactions on Internet Technology (TOIT), 3 (2003), pp. 1–27.

[19] A. Farahat, M. Bailey, *How Effective is Targeted Advertising?*, Proceedings of the 21st World Wide web Conference 2012, Lyon, France, April 16-20, 2012, pp. 111-120

[20] G. Gan, M. Chaoqun, W. Jianhong, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, Alexandria, 2007

[21] B. Ganter, R. Wille, *Formal Concept Analysis. Mathematical Foundations.* Springer, Berlin-Heidelberg-New York(1999).

[22] B. Goncalves, J. Ramasco, *Human dynamics revealed through Web analytics*, Phys. Rev. E 78, 026123 (2008).

[23] Z. Gyongy, H. Garcia-Molina, P. Berkhin, J. Pedersen, *Link Spam Detection Based on Mass Estimation*, 32nd International Conference in Very Large Data Bases (VLDB), Seoul, Korea, 2006, pp. 439-450

[24] **D. Haliţă**, D. Bufnea: *A study regarding inter domain linked documents similarity and their consequent bouncerate*, Studia Universitatis Babeş-Bolyai Informatica, 1/2014, pp. 83-91

[25] A. Huang, *Similarity Measures for Text Document Clustering*, Proceedings od the New Zealand Computer Science Research Student Conference, Hamilton, New Zealand, 2008, pp. 49-56

[26] R. Jaeschke, A. Hotho, C. Schmitz, B.Ganter, G. Stumme, *Trias - An Algorithm for Mining Iceberg Trilattices*, Proceedings of the IEEE International Conference on Data Mining, pp. 907-911, Hong Kong, IEEE Computer Society, 2006.

[27] The Joint Industry Committee for Web Standards (JICWEBS), *Reporting standards. website traffic.* Auditing Bureau of Circulations electronic (ABCe), Report 1, 2011.

[28] R. Kosala, H. Blockeel, *Web mining research: A survey*, ACM Sigkdd Explorations Newsletter, 2 (2000), pp. 1–15.

[29] F. Lehmann, R. Wille, *A Triadic Approach to Formal Concept Analysis*, Conceptual Structures: Applications, Implementation and Theory, vol. 954 of Lecture Notes in Artificial Intelligence, Springer Verlag, 1995.

[30] J. Leskovec, A. Rajaraman, J. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2010

[31] M. Najork, *Detecting Spam web Pages through Content Analysis*, International World Wide web Conference Comittee, Edinburgh, Scotland, 2006, pp. 83-92

[32] J.P. Norguet, B. Tshibasu-Kabeya, G. Bontempi, E. Zimanyi, *A Page-Classification Approach to Web Usage Semantic Analysis*, Engineering Letters, 14:1, EL 14 1 21.

[33] L. Rao, *WordPress Now Powers 22 Percent Of New Active websites In The U.S.*, August, 2011, TechCrunch

[34] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, S. Ventura, *Web usage mining for predicting final marks of students that use moodle courses*, Computer Applications in Engineering Education, 21 (2013), pp. 135–146.

[35] C. Romero, S. Ventura, A. Zafra, P. D. Bra, *Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems*, Computers & Education, 53 (2009), pp. 828–840.

[36] A. Singhal, *Modern Information Retrieval: A Brief Overview*, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2011, 24 (4): 35-43

[37] M. Spiliopoulou, L. C. Faulstich, *Wum: a tool for web utilization analysis*, in The World Wide Web and Databases, Springer, 1999, pp. 184–203.

[38] N. Spirin, J. Han, *Survey on web Spam Detection: Principles and Algorithms*, ACM SIGKDD Explorations Newsletter, Volume 13, Issue 2, December 2011, pp. 50-64

[39] J. Srivastava, R. Cooley, M. Deshpande, T. Pang-Ning, *Web usage mining: Discovery and applications of usage patterns from web data.*, SIGKDD Explorations, 1(2), 2000.

[40] R. Wille, *Conceptual landscapes of knowledge: a pragmatic paradigm for knowledge processing*, Proceedings of the International Symposium on Knowledge Representation, Use, and Storage Efficiency. Simon Fraser University, Vancouver 1997, 2-13.

[41] D.J. Watts, *Nature 445*, 489 (2007).

[42] Wille, R.: Conceptual Landscapes of Knowledge: a Pragmatic Paradigm for Knowledge Processing, In: Gaul, W.; Locarek-Junge H. (Eds.): Classification in the Information Age, Proceedings of the 22nd Annual Gfki Conference, Dresden, March 4-6, 1998, pp. 344–356.

[43] R. Wille, *Methods of Conceptual Knowledge Processing, Formal Concept Analysis*, 4th International Conference ICFCA 2006, Dresden, Germany, LNAI 3874, Springer 2006, 1–29.

[44] R. Wille, *The Basic Theorem of triadic concept analysis*, Order, no. 2, 1995,pp. 149-158

[45] Wordpress, 404 Redirected Wordpress *plugin*, *http://wordpress.org/ extend/plugins/404-redirected/*

[46] B. Zhou, S. C. Hui, K. Chang, *A formal concept analysis approach for web usage mining*, in Intelligent information processing II, Springer, 2005, pp. 437–441.

[47] D. Zhou, C. Burges, T. Tao, *Transductive Link Spam Detection*, Proceedings of the *3rd* International Workshop on Adversarial Information Retrieval on the web, AIRweb, New York, USA, ACM Press, 2007, pp. 21-28

[48] B. Zhou, S. C. Hui, A. C. Fong, *Web usage mining for semantic web personalization*, in Workshop on Personalization on the Semantic Web, 2005, pp. 66–72.

[49] Zyxware Technologies, Search404: Automatically search for content when a 404 error occurs, *http://drupal.org/project/search404*