

Universitatea "Babeş-Bolyai", Cluj-Napoca, România
Facultatea de Matematică și Informatică

Relevanța conținutului Web și a comportamentului utilizatorilor în analiza traficului

Teză de doctorat

Student doctorand:
Diana-Florina HALIȚĂ (căs. ȘOTROPA)

Coordonator științific:
Prof. Dr. Florian Mircea BOIAN

2017

Rezumat

Teza de doctorat "Relevanța conținutului *Web* și a comportamentului utilizatorilor în analiza traficului" oferă o imagine de ansamblu asupra metodelor cantitative și calitative de vizualizare, navigare și explorare a datelor, împreună cu studii de caz care dovedesc eficiența acestora, considerând seturi de date diferite.

Navigarea pe *Internet* a devenit un aspect esențial din viața de zi cu zi a fiecăruia dintre noi. *Web*-ul permite publicarea informațiilor online, precum și citirea, ascultarea și vizualizarea acestora. 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 cunoștințe. Printre aplicațiile WWW-ului se regăsesc: prezentarea în format electronic a publicațiilor, prezentarea produselor companiilor prin intermediul reclamelor, tranzacții comerciale, educație interactivă. Dezvoltarea rapidă a *World Wide Web*, precum și varietatea resurselor care sunt disponibile în *Internet*, impun necesitatea existenței unor instrumente care pot extrage informații importante din resursele existente sau care pot duce la îmbunătățirea calității experiențelor utilizatorilor în timpul explorării acestora. Astfel de instrumente joacă un rol important atât pentru proprietarii site-urilor *Web*, cât și pentru utilizatorii obișnuiți. Din perspectiva unor experți interesul major ar fi capturarea și analizarea traficului *Web*, cu scopul de a deservi interesul utilizatorilor. În acest context, lucrarea de față se concentrează pe prezentarea a două abordări diferite.

Prima abordare se referă la descoperirea de informații importante disponibile pe *Web*. Cea de-a doua se referă la urmărirea și analiza comportamentului utilizatorilor și a modalităților în care acesta este influențat de conținutul care se regăsește online. De ce aceste două abordări? Descoperirea de informații importante pe *Web* reprezintă procesul de căutare și analizare a resurselor disponibile online. Analiza conținutului ca tehnică de cercetare, este obiectivă, sistematică și poate defini documentul atât din punct de vedere cantitativ, cât și din punct de vedere calitativ.

Privind prima abordare urmărită în această lucrare, luând în considerare analiza unui document care este interesant sau util, se poate pune problema găsirii unor 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 confluența mai multor clase: analiza legăturilor dintre documente și analiza

consecințelor ce pot apărea în urma acestor legături.

Privitor la cea de-a doua abordare, scopul acestei teze este acela de a găsi noi modalități 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 și din punct de vedere al tipului de conținut prezentat).

Majoritatea cercetătorilor propun utilizarea tehnicilor din *Web Mining* pentru a putea studia astfel de comportamente. *Web Mining* descrie:

- procesul de extragere a informațiilor din resursele disponibile online, atât prin analizarea conținutului cât și a prezentării;
- procesul de deducere a informațiilor din structura de *link*-uri a unui site *Web*;
- procesul de analizare a informațiilor referitoare la vizitele paginilor *Web* salvate în fișierele de log ale serverelor *Web*, în vederea descoperirii de modele de navigare interesante necunoscute anterior și posibil utile.

Abordarea folosită în teză propune utilizarea Analizei Conceptuale Formale - (FCA) ca teorie matematică cu aplicații în procesarea conceptuală a datelor. Aceasta se ocupă cu evidențierea corelațiilor din seturile de date împreună cu motivarea existenței unor astfel de conexiuni.

Analiza Conceptuală Formală face parte din domeniul cunoscut sub numele de procesarea conceptuală a datelor. FCA se ocupă cu înțelegerea faptelor și a corelațiilor într-un set de date și totodată încearcă să ofere explicații pentru existența acestor fapte și legături între date. Analiza Conceptuală Formală Poliadică este o extensie a cazului clasic care, în loc de relații binare, se bazează pe relații de incidență n -are pentru a defini concepte formale, adică clusterse maximale de date în care toate elementele sunt corelate prin relația de incidență. În această lucrare este prezentată o imagine de ansamblu asupra metodelor de vizualizare și explorare pentru contexte formale poliadice definite ca seturi poliadice de date împreună cu o relație de incidență n -ară.

Cuvintele cheie pentru temele abordate în această teză sunt: migrarea conținutului *Web*, sistem de management de conținut, rată de respingere, similaritatea documentelor *Web*, platformă educațională, Analiză Conceptuală Formală Poliadică, Analiză Conceptuală Formală Temporală, *Web Usage Behavior*, tipare de navigare, traiectorie, inițiator de comportament.

Teza are următoarea structură:

1 Introducere

1.1 Introducere

1.2 Motivație

-
- 1.3 Scopul lucrării
 - 1.4 Contribuții personale
 - 2 Preliminarii
 - 2.1 Caracterizarea traficului Web
 - 2.2 Server proxy: Squid
 - 2.3 Măsuri de similaritate
 - 2.3.1 Măsura de similaritate Cosinus
 - 2.3.2 Măsura de similaritate Jaccard
 - 2.3.3 Măsura de similaritate Sorensen
 - 2.3.4 Măsura de similaritate Jaro-Winkler
 - 2.4 Web Mining
 - 2.4.1 Web Content Mining
 - 2.4.2 Web Structure Mining
 - 2.4.3 Web Usage Mining
 - 2.5 Analiză Conceptuală Formală
 - 2.5.1 Analiza Conceptuală Formală Diadică
 - 2.5.1.1 Scala nominală
 - 2.5.1.2 Scala ordinală
 - 2.5.1.3 Scala interordinală
 - 2.5.2 Analiza Conceptuală Formală Triadică
 - 2.5.3 Analiza Conceptuală Formală Poliadică
 - 2.5.4 Analiza Conceptuală Formală Temporală
 - 2.6 Stadiul actual al cunoașterii în domeniu
 - 3 Relevanța conținutului Web și a comportamentului utilizatorilor în analiza traficului
 - 3.1 Impactul similarității documentelor Web asupra traficului
 - 3.1.1 Migrarea transparentă a unui site Web între două sisteme de management de conținut
 - 3.1.1.1 Provocările procesului de migrare
 - 3.1.1.2 Tipuri de migrare a conținutului
 - 3.1.1.3 Algoritm și implementare
 - 3.1.1.4 Rezultatele obținute și evaluarea acestora

-
- 3.1.2 Analizarea legăturii dintre similaritatea documentelor Web și a ratei de respingere generate de link-urile dintre aceste documente
 - 3.1.2.1 Rezultate anterioare
 - 3.1.2.2 Estimarea ratei de respingere folosind similaritatea conținutului dintre sursă și destinație
 - 3.1.3 Concluzii și direcții de cercetare
 - 3.2 Determinarea caracteristicilor site-urilor Web de tip spam
 - 3.2.1 Măsurarea și vizualizarea nivelului de scrappiness al unui site Web
 - 3.2.1.1 Preliminarii
 - 3.2.1.2 Metodologia de cercetare, experimente și rezultate
 - 3.2.1.3 Validarea rezultatelor
 - 3.2.2 Concluzii și direcții de cercetare
 - 4 Dinamica utilizatorilor în site-uri educaționale: aspecte cantitative și calitative
 - 4.1 Interpretarea log-urilor unei platforme de e-learning folosind Analiza Conceptuală Formală
 - 4.1.1 Metrici de analiză a datelor
 - 4.1.2 Web Usage Mining pe platforme educaționale
 - 4.1.3 Platforma de e-learning: PULSE
 - 4.1.4 Vizualizarea într-un format circular a datelor: CIRCOS
 - 4.1.5 ToscanaJ - instrumentul de analiză calitativă a datelor
 - 4.1.6 Analiza comportamentului utilizatorilor Web folosind Analiza Conceptuală Formală
 - 4.1.7 Analizarea efectului schimbării metodelor de predare și evaluare folosind Analiza Conceptuală Formală
 - 4.1.8 Concluzii și direcții de cercetare
 - 4.2 Tipare comportamentale în utilizarea platformelor de e-learning
 - 4.2.1 Atractori în sistemele educaționale
 - 4.2.1.1 Investigarea tiparelor comportamentale prin procesarea conceptuală a datelor
 - 4.2.2.1 Atractori educaționali
 - 4.2.2.2 Atractori populari: ramificarea lanțurilor de pagini vizitate
 - 4.2.2.3 Atractori critici
 - 4.2.2 Concluzii și direcții viitoare de cercetare
 - 4.3 Evidențierea structurilor conceptuale din log-urile Web folosind Analiza Conceptuală Formală Poliadică

-
- 4.3.1 Investigarea grupurilor de comportamente repetitive în platforme de e-learning folosind Analiza Conceptuală Formală Poliadică
 - 4.3.1.1 Analiza Formală Conceptuală Tetrică
 - 4.3.2 Investigarea inițiatorilor de comportamente în platforme de e-learning folosind Analiză Formală Conceptuală Poliadică și Programarea cu mulțimi de răspuns
 - 4.3.3 Concluzii și direcții viitoare de cercetare
 - 4.4 Evidențierea structurilor conceptuale din log-urile Web folosind Analiza Conceptuală Formală Temporală
 - 4.4.1 Investigarea atractorilor educaționali și a tiparelor de navigare în timp
 - 4.4.2 Investigarea comportamentului utilizatorilor în platformele educaționale folosind Analiza Conceptuală Formală Temporală
 - 4.4.2.1 Atractori de navigare
 - 4.4.2.2 Atractori populari
 - 4.4.2.3 Atractori critici
 - 4.4.2.4 Discuție: traiectoriile studenților
 - 4.4.2.5 Exemplu: formalizarea traiectoriei unui student
 - 4.4.3 Dinamica temporală a utilizatorilor
 - 4.4.4 Concluzii și direcții de cercetare
 - 5 Concluzii și direcții viitoare de cercetare
 - 6 Anexe
 - Lista algoritmilor
 - Lista figurilor
 - Lista tabelor
 - Acronime
 - Glosar
 - Lista de publicații

După cum reiese din structura prezentată anterior, părțile trei și patru evidențiază contribuțiile principale ale tezei. Aceste contribuții constau în definirea unor metode noi de:

- direcționare a utilizatorilor care au fost direcționați greșit de motoarele de căutare;

-
- de analiză a legăturii dintre documentele *Web* și rata de respingere generată de *link*-urile dintre aceste documente;
 - de identificare a site-urilor *Web* de tip spam;
 - de navigare și explorare în seturi n-adice de date;
 - de identificare a tiparelor comportamentale ale utilizatorilor unei platforme educaționale;
 - de investigare și evidențiere a structurilor conceptuale din seturi de date folosind Analiza Conceptuală Formală Poliadică și Temporală.

În cele ce urmează sunt descrise pe scurt capitolele cuprinse în partea a treia.

În Capitolul 3.1 este descrisă modalitatea în care similaritatea documentelor regăsite online își pune amprenta asupra traficului *Web*. Capitolul începe cu o secțiune în care este prezentată o metodă de migrare transparentă a unui site *Web*, atât din perspectiva utilizatorilor, cât și din perspectiva motoarelor de căutare. Totodată este propusă și o metodă de estimare a ratei de respingere pentru terțe site-uri cu posibile aplicații în ierarhizarea site-urilor *Web*.

În Capitolul 3.2 sunt prezentate modalitățile prin care site-urile *Web* de tip spam pot fi identificate. Astfel, motoarele de căutare pot fi asistate în detectarea acestor tipuri de site-uri, care datorită conținutului pe care îl prezintă, ar putea fi penalizate sau chiar eliminate din SERP (*Search Engine Results Page*).

În cele ce urmează sunt descrise pe scurt capitolele cuprinse în partea a patra.

În Capitolul 4.1 este analizat comportamentul studenților pe platforma educațională PULSE. În acest capitol sunt aplicate tehnicile din *Web Usage Mining* într-un context educațional și sunt prezentate metricile de analiză a datelor. De asemenea, este oferită o explicație a faptului că tehnicile obișnuite de *Web Usage Mining* nu dau rezultate bune în cazul unui sistem de e-learning. În consecință se propune Analiza Conceptuală Formală ca tehnică de *Web Usage Mining* care se folosește pentru analiza detaliată a log-urilor platformei PULSE. În acest sens, sunt descrise în detaliu cele trei etape ale analizei: pre-procesarea datelor, descoperirea și analiza tiparelor comportamentale. În procesul de descoperire a tiparelor comportamentale se investighează comportamentul studenților în funcție de diferite aspecte temporale ale cursului. În final, aceste tipare sunt vizualizate într-un format circular folosind CIRCOS. În interpretarea rezultatelor se identifică corelații între comportamentul studenților și activitățile din timpul cursului, cum ar fi atribuirea temelor de laborator, examen parțial, examen final.

În Capitolul 4.2 sunt investigate tiparele comportamentale generate de vizitatori odată cu utilizarea platformelor de e-learning. Au fost definiți atractorii, care sunt tipare comportamentale la care utilizatorii aderă în timp ce utilizează platforma educațională. Unii atractori sunt sugerați sau doriți de către educator sau

sunt impuși de structura sistemului educațional. Aceștia oferă instructorului o privire de ansamblu asupra modalității în care utilizatorii navighează în platforma educațională.

În Capitolul 4.3 sunt prezentate modalitățile prin care se pot evidenția structurile conceptuale din log-urile *Web* generate de vizitatori odată cu navigarea prin platforma educațională. Totodată este prezentată noțiunea de inițiator de comportamente, adică utilizator care a aderat prima dată la un anumit comportament și care generează o mulțime de utilizatori care urmează același comportament. Tot aici sunt prezentate și modalitățile în care alți utilizatori asimilează comportamentele inițiatorilor de comportamente și analiza comportamentelor repetitive în cadrul unui grup de utilizatori.

În Capitolul 4.4 sunt prezentate modalitățile prin care se pot evidenția structurile conceptuale din log-urile *Web* generate de vizitatori odată cu utilizarea platformelor de e-learning folosind Analiza Formală Conceptuală Temporală. De asemenea este prezentată evoluția temporală a tiparelor comportamentale generate de un grup de utilizatori, demonstrând astfel eficiența combinării tehnicii construirii de scale conceptuale cu metode ale Analizei Conceptuale Formale Temporale. Totodată sunt definite formal noțiunile introduse empiric în celelalte capitole.

În concluziile tezei este evidențiată importanța contribuțiilor prezentate și sunt sugerate viitoarele direcții de cercetare. De asemenea, în viitor, vom continua să dezvoltăm și să îmbunătățim metodele de detectare (semi) automată a site-urilor care prezintă informație de calitate îndoielnică, dar care prin modul de prezentare urmăresc creșterea numărului de vizite și respectiv scăderea ratei de respingere a site-ului.