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

Relevanţa conţinutului Web şi a comportamentului utilizatorilor în analiza traficului

Teză de doctorat

Student doctorand:
Diana-Florina HALIŢĂ (căs. ŞOTROPA)
Coordonator ştiinţific:
Prof. Dr. Florian Mircea BOIAN

2017

Rezumat

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

Navigarea pe *Internet* a devenit un aspect esenţial din viaţa de zi cu zi a fiecăruia dintre noi. *Web*-ul permite publicarea informaţiilor online, precum şi citirea, ascultarea şi 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 cunoştinţe. Printre aplicaţiile WWW-ului se regăsesc: prezentarea în format electronic a publicaţiilor, prezentarea produselor companiilor prin intermediul reclamelor, tranzacţii comerciale, educaţie interactivă. 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 proprietarii site-urilor *Web*, cât 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? Descoperirea de informaţii importante pe *Web* reprezintă procesul de căutare şi analizare a resurselor disponibile online. 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 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 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 teze 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 Mining* pentru a putea studia astfel de comportamente. *Web Mining* descrie:

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

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

Analiza Conceptuală Formală face parte din domeniul cunoscut sub numele de procesarea conceptuală a datelor. FCA se ocupă cu înţelegerea faptelor şi a corelaţiilor într-un set de date şi totodată încearcă să ofere explicaţii pentru existenţa acestor fapte şi legături între date. Analiza Conceptulă Formală Poliadică este o extensie a cazului clasic care, în loc de relaţii binare, se bazează pe relaţii de incidenţă n-are pentru a defini concepte formale, adică clustere maximale de date în care toate elementele sunt corelate prin relaţia de incidenţă. În această lucrare este prezentată o imagine de ansamblu asupra metodelor de vizualizare şi explorare pentru contexte formale poliadice definite ca seturi poliadice de date împreună cu o relaţie de incidenţă n-ară.

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

Teza are următoarea structură:

După cum reiese din structura prezentată anterior, părţile trei şi patru evidenţiază contribuţiile principale ale tezei. Aceste contribuţii constau în definirea unor metode noi de:

- direcţionare a utilizatorilor care au fost direcţionaţi greşit de motoarele de căutare;

- de analiză a legăturii dintre documentele *Web* şi rata de respingere generată de *link*-urile dintre aceste documente;
- de identificare a site-urilor *Web* de tip spam;
- de navigare şi explorare în seturi n-adice de date;
- de identificare a tiparelor comportamentale ale utilizatorilor unei platforme educaţionale;
- de investigare şi evidenţiere a structurilor conceptuale din seturi de date folosind Analiza Conceptuală Formală Poliadică şi 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 îşi pune amprenta asupra traficului *Web*. Capitolul începe cu o secţiune în care este prezentată o metodă de migrare transparentă a unui site *Web*, atât din perspectiva utilizatorilor, cât şi din perspectiva motoarelor de căutare. Totodată este propusă şi o metodă de estimare a ratei de respingere pentru terţe site-uri cu posibile aplicaţii în ierarhizarea site-urilor *Web*.

În Capitolul 3.2 sunt prezentate modalităţile 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 conţinutului 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 comportamentului studenţilor pe platforma educaţională PULSE. În acest capitol sunt aplicate tehnicile din *Web Usage Mining* într-un context educaţional şi sunt prezentate metricile de analiză a datelor. De asemenea, este oferită o explicaţie a faptului că tehnicile obişnuite de *Web Usage Mining* nu dau rezultate bune în cazul unui sistem de e-learning. În consecinţă se propune Analiza Conceptuală Formală ca tehnică de *Web Usage Mining* care se foloseşte pentru analiza detaliată a log-urilor platformei PULSE. În acest sens, sunt descrise în detaliu cele trei etape ale analizei: pre-procesarea datelor, descoperirea şi analiza tiparelor comportamentale. În procesul de descoperire a tiparelor comportamentale se investighează comportamentul studenţilor în funcţie de diferite aspecte temporale ale cursului. În final, aceste tipare sunt vizualizate într-un format circular folosind CIRCOS. În interpretarea rezultatelor se identifică corelaţii între comportamentul studenţilor şi activităţile din timpul cursului, cum ar fi atribuirea temelor de laborator, examen parţial, examen final.

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

sunt impuşi de structura sistemului educaţional. Aceştia oferă instructorului o privire de ansamblu asupra modalităţii în care utilizatorii navighează în platforma educaţională.

În Capitolul 4.3 sunt prezentate modalităţile prin care se pot evidenţia structurile conceptuale din log-urile *Web* generate de vizitatori odată cu navigarea prin platforma educaţională. Totodată este prezentată noţiunea de iniţiator de comportamente, adică utilizator care a aderat prima dată la un anumit comportament şi care generează o mulţime de utilizatori care urmează acelaşi comportament. Tot aici sunt prezentate şi modalităţile în care alţi utilizatori asimilează comportamentele iniţiatorilor de comportamente şi analiza comportamentelor repetitive în cadrul unui grup de utilizatori.

În Capitolul 4.4 sunt prezentate modalităţile prin care se pot evidenţia 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 evoluţia temporală a tiparelor comportamentale generate de un grup de utilizatori, demonstrând astfel eficienţa combinării tehnicii construirii de scale conceptuale cu metode ale Analizei Conceptuale Formale Temporale. Totodată sunt definite formal noţiunile introduse empiric în celelalte capitole.

În concluziile tezei este evidenţiată importanţa contribuţilor prezentate şi sunt sugerate viitoarele direcţii de cercetare. De asemenea, în viitor, vom continua să dezvoltăm şi să îmbunătăţim metodele de detectare (semi) automată a site-urilor care prezintă informaţie de calitate îndoielnică, dar care prin modul de prezentare urmăresc creşterea numărului de vizite şi respectiv scăderea ratei de respingere a site-ului.