Rezumat

La începutul anilor 2000 web-ul avea peste 800 de milioane de situri care acopereau cele mai diverse domenii şi avea 6 Tocteți de date memorate pe aproximativ trei milioane de servere. Zilnic aproape un milion de pagini sunt adăugate şi în medie o pagină este modificată la câteva luni, iar într-o lună sunt modificate câteva sute de Gocteți. Dezvoltarea rapidă a acestor caracteristici face ca Web-ul să devină o oportunitate şi o mare provocare pentru cercetătorii din domeniul ştiinței calculatoarelor.

O falsă soluție sugerată de unii organizatori are fi organizarea informațiilor pe măsură ce ele sunt generate, în conformitate cu nişte reguli prestabilite. Acest lucru este inutil, deoarece majoritatea utilizatorilor nu se vor obosi să le respecte. Orice încercare de aplicare strictă a unor reguli de organizare va determina utilizatorii să plece. Rezultă că pentru acest moment majoritatea informațiilor vor fi organizate după ce au fost generate, iar instrumentele de căutare vor trebui să lucreze în strânsă cooperare cu instrumentele de organizare.

În acest referat încerc să prezint o scurtă perspectivă asupra cercetării în acest domeniu fertil şi tendințele acesteia. Deoarece cantitatea de informații disponibilă pe web creşte continuu, volumul datelor nestructurate (text şi hipertext) depăşeşte tot mai mult volumul de date structurate (bazele de date). Astfel, cercetarea din domeniul tehnicilor de învățare automată se îndreaptă tot mai mult spre dezvoltarea algoritmilor pentru analiza datelor nestructurate. Pentru a găsi o cale de a face căutarea informațiilor pe web mai uşoară în ultimii ani cercetarea în acest domeniu se îndreaptă în principal spre 4 domenii: strategie, algoritmi, arhitectură şi interfața cu utilizatorul.

Pentru utilizatorul care vrea să găsească ceva pe web, căutarea poate deveni o activitate foarte dificilă în momentul în care motorul de căutare îi oferă mii de documente pentru o interogare dată. Astfel multe cercetări sunt concentrate pe organizarea rezultatelor căutării, găsirea unor noi strategii de implementare a motoarelor de căutare şi în arhitecturi de reprezentare a rezultatelor căutării, luând mai puțin în calcul utilizatorul. În ultimii ani tot mai multe persoane ne-experimentate şi din diferite domenii de activitate accesează şi chiar modifică conținutul web-ului. Astfel o nouă provocare este aceea de a face web-ul cât mai simplu de utilizat şi cât mai pe înțelesul tuturor. Din acest punct de vedere, cercetarea este împărțită pe trei direcții: (1) găsirea informațiilor relevante, (2) reprezentarea rezultatelor căutării şi (3) ajutarea utilizatorului în găsirea informațiilor.

Majoritatea motoarelor de căutare actuale *găsesc informații relevante* (informațiile care satisfac nevoile utilizatorului) pentru cuvintele introduse de utilizator în interogare utilizând metode statice (cum ar fi PageRanks) care iau în considerare doar informații despre structura şi conținutul web-ului, fără să ia în considerare informații despre utilizatorul sau utilizatorii care sunt interesați de acele informații. Provocarea în acest domeniu este construirea profilului utilizatorului. S-au prezentat metode care construiesc profilul utilizatorului static, utilizând ceea ce introduce utilizatorul despre el şi câteva documente inițiale pre-etichetate de utilizator. Problema este că în timp utilizatorul nu îşi reactualizează profilul, iar de obicei şi interesul utilizatorului într-un anumit domeniu se modifică destul de frecvent. Ca o soluție la această problemă se încearcă crearea şi modificarea profilului utilizator dinamic, în timp ce acesta caută pe web. Pentru aceasta pot fi folosite noile documente găsite pe care acesta le consideră interesante (în concordanță cu cea ce îl interesează pe utilizator).

Studiile asupra *reprezentării rezultatelor* au arătat că atât din punct de vedere obiectiv cât şi din punct de vedere subiectiv o intefață în care rezultatele sunt grupate după categorii este superioară

unei interfețe în care rezultatele sunt ordonate după „gradul de potrivire" cu interogarea. O provocare în acest domeniu este generalizarea rezultatelor în domenii, cu alte cuvinte cum am putea crea domeniile pe baza rezultatelor căutării astfel încât să fie cât mai sugestive şi mai intuitive pentru utilizator. O altă provocare ar fi reprezentarea acestora într-un mod cât mai uşor de înțeles pentru utilizator. Astfel, se încearcă diferite tipuri de reprezentări a rezultatelor căutării cum ar fi o reprezentare ierarhică sau o reprezentare grafică 3D. Din acest punct de vedere reprezentarea ierarhică este mult mai intuitivă, mai uşor de utilizat şi de înțeles pentru utilizator. Problema principală a acesteia este crearea şi organizarea categoriilor. Dacă se merge pe ideea creării unei structuri fixe de categorii (intuitivă şi uşor de folosit pentru utilizator) apare problema că toate rezultatele căutării trebuie să fie grupate în acea structură fixă de categorii. O altă abordare a fost crearea unor structuri dinamice de categorii, în funcție de rezultatele căutării. Problema este selectarea caracteristicilor care sunt într-adevăr relevante pentru utilizator şi care vor fi utilizate pentru clasificarea documentelor. O altă problemă a acestui tip de reprezentare este că doar o mică porțiune din cele mai reprezentative şi importante informații sunt afişate în fereastra inițială şi astfel se încearcă folosirea tehnicilor de suprapunere pentru a acoperi cât mai multe detalii. Pentru a rezolva această problemă se încearcă reprezentarea rezultatelor căutării într-o perspectivă 3D astfel ca mult mai multe informații să poată fi prezentate simultan utilizatorului şi să se poată prezenta şi corelațiile care există între aceste rezultate. Vizualizarea şi explorarea rezultatelor căutării într-o asemenea perspectivă constituie o mare problemă pentru creatorii de interfețe (web designer) deoarece ei trebuie să grupeze cât mai bine rezultatele căutării astfel ca utilizatorul să poată înțelege şi să furnizeze o metodă de interacțiune eficientă pentru explorare.

O altă abordare este încercarea de a *ajuta utilizatorul în găsirea informațiilor* prezentându-se acestuia diferite sugestii pentru îngustarea domeniului de căutare, în funcție de cuvintele de interogare. În majoritatea cazurilor utilizatorul nu ştie exact cum să caute ceea ce doreşte. Astfel anumite sisteme de căutare încearcă să prezinte diferite sinonime la cuvintele de căutare specificate de utilizator sau diferite domenii distincte în care pot apărea acele cuvinte. Se îngustează astfel din start domeniul de căutare. O altă abordare este încercarea de a extrage din documentele inițiale găsite de utilizator şi considerate interesante a anumitor cuvinte cheie, care după aceea să poată fi folosite pentru găsirea altor documente relevante. Unele sisteme de căutare, tot pentru a veni în sprijinul utilizatorilor, încearcă să grupeze mai mulți utilizatori în funcție de domeniul lor de interes şi de ceea ce consideră un utilizator că este relevant pentru el. Când apare un nou utilizator se încearcă încadrarea acestuia în anumite grupuri în funcție de ceea ce îl interesează ca apoi să i se poată sugera domenii (documente) găsite de ceilalți utilizatori din acel grup ca fiind interesante din acel punct de vedere, astfel încercând să deschidă noi perspective utilizatorului.

Perspectiva prezentată în acest referat face parte dintr-un domeniu de cercetare mult mai vast în care se încearcă găsirea unei mai bune reprezentări şi organizări a informațiilor de pe web, sporind comunicația şi colaborarea dintre utilizatori şi aplicație. Perspectiva prezentată în acest referat este orientată pe încercarea de a reorganiza dinamic interfața web-ului din punct de vedere al utilizatorului, fără modificarea structurii de bază. Astfel cercetătorii încearcă să personalizeze interfața web-ului pentru fiecare utilizator în parte, făcând-o mult mai uşoară şi accesibilă. În abordarea globală se încearcă transformarea orientări web-ului de la orientarea pe documente la cea pe date relevante, adică de la orientarea spre om la orientarea spre maşină. Cercetătorii din acest domeniu încearcă să reorganizeze datele de pe web atribuindu-le un înțeles, făcând astfel posibilă procesarea automată a acestora. Această nouă reprezentare este numită „Web Semantic" şi vrea să organizeze datele şi informațiile într-un mod natural pentru utilizator şi în acelaşi timp uşor de utilizat pentru procesarea automată. Se poate considera că două programe pun împreună cunoştințele lor schimbând ontologii care furnizează vocabularul necesar pentru discuție. Această

idee nu este în totalitate uşor de implementat deocamdată datorită dimensiunii web-ului şi datorită dificultății de a crea ideea generală de reprezentare a acestuia, luând în considerare contextul.

Acest referat este structurat în două părți principale (capitolele 2 şi 3). Capitolul 2 cuprinde o prezentare mai în detaliu a mineritului datelor care a fost inclusă pentru clarificarea conceptelor din acest domeniu, utilizate şi în procesul de minerit al fişierelor text. Apoi sunt prezentați algoritmii de minerit al fişierelor text şi modificările aduse algoritmilor folosiți în mineritul datelor. Aceştia sunt necesari în dezvoltarea următorului referat. În încheierea acestui capitol este prezentată partea de minerit a conținutului web-ului şi a utilizării acestuia pentru a obține cea ce ne interesează în final şi anume crearea profilului utilizatorului. În capitolul 3 este prezentată abordare actuală asupra reorgnizării rezultatelor căutării oferite de motoarele de căutare actuale. În acest capitol prezentăm atât abordările deja implementate cât şi cercetările în curs.

Mulțumiri

Țin să exprim mulțumirile mele către conducătorul meu de doctorat prof. univ. dr. ing. Lucian VINȚAN, căruia îi sunt recunoscător în primul rând pentru responsabilitatea cu care dânsul îşi priveşte meseria, apoi pentru discuțiile extrem de stimulative pe care le-am avut şi pentru sprijinul acordat. De asemenea, rămân îndatorat pentru sprijinul profesional competent acordat mie cu generozitate încă de la începuturile perioadei de pregătire prin doctorat de către profesorii: mat. Ioana MOISIL, ing. Boldur BĂRBAT, ing. Daniel VOLOVICI, ing. Dorin SIMA şi ing. Macarie BREAZU.

Aş vrea să mulțumesc firmei SIEMENS AG, Corporate Technology, Information & Communication Division, MUNCHEN Germania, şi în special vicepreşedintelui Dr.h.c. Hartmut RAFFLER, pentru sponsorizarea acordată acestui program de doctorat. Deosebit de utile mi-au fost sugestiile profesionale precum şi suportul material oferit. De asemenea vreau să-i mulțumesc tutorelui meu de la firma SIEMENS, dr. Volker TRESP, Senior Principal Research Scientist in Neural Computation, pentru sprijinul oferit şi pentru îndrumarea în acest imens şi dificil domeniu al cercetării. Aş vrea de asemenea să-i mulțumesc d-lui Dr. Kai YU de la aceeaşi companie pentru informațiile furnizate în dezvoltarea ideilor. Şi nu în ultimul rând aş vrea să mulțumesc tuturor celor care m-au ajutat în realizarea acestui referat.