

Îmbunătățirea relevanței rezultatelor motoarelor de căutare folosind informații semantice din Wikipedia

Cristina Șcheau

Universitatea
„Politehnica” din
București,
Splaiul
Independenței, Nr
313 București
cristina.scheau@
cti.pub.ro

Traian Rebedea

Universitatea
„Politehnica” din
București,
Splaiul
Independenței, Nr
313 București
traian.rebedea@
cs.pub.ro

Costin Chiru

Universitatea
„Politehnica” din
București,
Splaiul
Independenței, Nr
313 București
costin.chiru@
cs.pub.ro

Ștefan Trăușan-Matu

Universitatea „Politehnica”
din București,
Splaiul Independenței, Nr
313 București
Institutul de Cercetare în
Inteligența Artificială al
Academiei Române,
Calea 13 Septembrie, Nr
13 București
trausan@cs.pub.ro

REZUMAT

În funcție de intenția utilizatorului, interogările către un motor de căutare pot fi împărțite în trei categorii: tranzacționale, informaționale și navigaționale [1]. Pentru a satisface cele trei tipuri de căutări, motoarele de căutare din prezent folosesc în principiu algoritmi de analiză a legăturilor între pagini îmbunătățiți cu un factor care depinde de numărul de apariții și ordinea cuvintelor căutate. Pentru interogările tranzacționale și informaționale, relevanța ar putea fi îmbunătățită dacă s-ar folosi și informații semantice despre conceptele căutate. Wikipedia este un tezaur imens care are avantajul de a fi deja tradusă în multe limbi, cu o structură densă de legături interne putând fi folosită pentru extragerea de informații în diverse moduri. Lucrarea de față propune o modalitate de extragere a relațiilor semantice din Wikipedia, iar apoi folosirea acestora pentru a determina ordinea în care rezultatele întoarse de un motor de căutare sunt afișate utilizatorului.

Cuvinte cheie

Căutare web, motor de căutare, optimizare, relații semantice, Wikipedia.

Clasificare ACM

H.3.3 Information Search and Retrieval, H.5.4 Hypertext/Hypermedia, I.2.7 Natural Language Processing.

INTRODUCERE

Cercetarea în domeniul motoarelor de căutare este destul de vastă și diferite idei au fost sugerate, atât din mediul academic cât și din industrie. În prezent, utilizatorii pot să obțină diferite rezultate în funcție de cuvintele cheie pe care le furnizează motorului de căutare. Rezultatele obținute sunt destul de relevante, însă se observă o lipsă a relațiilor semantice, ordonarea făcându-se după importanța paginii și conținutul lexical și nu după conținutul semantic. Mai mult, sunt multe informații redundante în aceste pagini, ceea ce mărește timpul de căutare al utilizatorului. În aceste condiții, această lucrare propune o

reordonare a rezultatelor aduse de un motor de căutare (de exemplu, Google), în funcție de conținutul semantic.

Pentru extragerea de informații semantice pot fi folosite două tipuri de baze de cunoștințe: baze de cunoștințe lingvistice (dicționare / ontologii precum WordNet) sau baze de cunoștințe colaborative precum Wikipedia. Folosind baze de cunoștințe lingvistice, au fost numeroase încercări de expansiune a cuvintelor cheie date de utilizator pentru a obține rezultate mai bune. Din păcate, conform experimentelor efectuate, această metodă nu a oferit rezultatele scontate. Principalele dezavantaje ale acestor instrumente lingvistice sunt puterea limitată de acoperire, probabilitatea de a fi neactualizate, dimensiunea fixă, restricționarea la un vocabular ce conține concepte generale în mare parte.

Bazele de cunoștințe colaborative acoperă aceste dezavantaje, oferind însă un text semi-structurat care necesită prelucrări ulterioare pentru a deveni cu adevărat o sursă de cunoștințe. Printre cele mai reprezentative astfel de baze de cunoștințe se numără Wikipedia și Wiktionary.

Wikipedia este o enciclopedie liberă, multilingvă și probabil una din cele mai mari colecții de informație disponibilă tuturor oamenilor. Versiunea în engleză numără peste 2,5 milioane de articole distincte (de fapt peste 6 milioane de pagini, incluzând și paginile de redirectare). Informația care poate fi prelucrată din această enciclopedie nu este numai textul propriu-zis al articolelor, ci și multitudinea de legături interne ale articolelor sau paginile de redirectare - acestea din urmă pot fi foarte utile la determinarea sinonimelor.

Metoda propusă pentru determinarea relațiilor semantice este următoarea: Pe baza legăturilor înainte și înapoi dintre articole, se determină distanța dintre articole. Pentru început, se consideră că articolul reprezintă conceptul din titlu, obținându-se astfel o relație între concepte. În momentul în care utilizatorul introduce cuvintele cheie, se folosește un motor de căutare pentru a obține cele mai importante rezultate. Pe cuvintele introduse de utilizator se aplică un proces de stemming, iar apoi se elimină cuvintele de stop. În paginile întoarse de motorul de căutare se caută și cuvintele aflate în relație cu cuvintele

cheie prelucrate. În funcție de numărul de apariții ale acestor concepte precum și de importanța lor se obține un nou scor care este combinat cu cel determinat de Google. Cuvintele cheie introduse de utilizator au importanță maximă, iar restul conceptelor au o importanță determinată în funcție de cât de puternică este legătura dintre acestea și un cuvânt cheie introdus de utilizator, folosind Wikipedia.

Restul lucrării este structurat în modul următor: în secțiunea următoare sunt prezentate alte metode de îmbunătățire a relevanței rezultatelor motoarelor de căutare precum și alte sisteme de extragere a informațiilor semantice folosind Wikipedia. În secțiunea 3 este descris în detaliu modelul propus, iar în secțiunea 4 sunt prezentate rezultatele obținute. În final, vor fi prezentate concluziile.

ABORDĂRI ANTERIOARE PENTRU ÎMBUNĂTĂȚIREA REZULTATELOR CĂUTĂRILOR WEB FOLOSIND INFORMAȚII SEMANTICE

Recalcularea ordinii rezultatelor unui motor de căutare web folosind WordNet

O primă lucrare interesantă în această direcție este Hu et al. [2], care prezintă în prima parte o metodă pentru a clasifica rezultatele obținute de la Google în categorii în funcție de similaritatea semantică. Programul preia ca date de intrare rezultatele aduse de Google pentru cuvintele cheie date de utilizator. Din aceste rezultate sunt extrase fragmente formate (snippets). Un snippet constă în conținutul text al unei pagini și titlul său. Fiecare snippet este pus într-o anumită categorie în funcție de similaritatea dintre cuvintele relevante din interiorul său și subiectul (topic-ul) categoriei respective.

Algoritmul de calcul al similarității semantice ales se bazează pe WordNet, calculând în principiu lungimea căii celei mai apropiate dintre cele două concepte pentru care se dorește determinarea distanței semantice (ținând cont de hiponime) precum și cantitatea de informație conținută de concept, care depinde de probabilitatea de apariție a conceptului în întreg tezaurul.

În a doua parte a lucrării se propune un nou algoritm pentru ordonarea rezultatelor. Astfel, noul scor se obține prin combinarea PageRank-ului [14] obținut de Google cu factorul de similaritate dintre snippets și topicul din interogare precum și cu un factor de timp. Pentru calculul factorului de timp se pornește de la premisa că utilizatorul dorește să vadă cea mai nouă informație. Astfel factorul este invers proporțional cu diferența dintre timpul curent și timpul de cache.

Algoritmul Topic-Sensitive PageRank

O altă abordare pornește de la ideea că este mai bine să se folosească un algoritm similar cu PageRank-ul clasic. În încercarea de a îmbunătăți algoritmul PageRank clasic cu informații semantice, s-a dezvoltat algoritmul Top-Sensitive PageRank [3]. Ideea acestui algoritm este de a avea în loc de un singur vector ca la PageRank, mai mulți vectori, fiecare pentru un topic anume. Utilizatorului îi sunt afișate rezultatele în funcție de vectorul topicului interogării introdus. Topicul interogării poate fi

determinat fie din context, fie prin calculul distanței față de toți vectorii, alegându-se cel cu distanța cea mai mică.

Utilizarea Wikipedia ca sursă de informații semantice

În ultima perioadă, ideea utilizării Wikipedia ca sursa de informații semantice este din ce în ce mai vehiculată. În primul rând, Wikipedia conține informații relevante pentru majoritatea domeniilor și conceptelor prezentate. În al doilea rând, numărul de concepte și domenii acoperite este foarte mare, depășind din punct de vedere al numărului de pagini orice alternativă. Poate că cel mai important avantaj este faptul că se folosește o structură de wiki, care îi conferă posibilitatea de a face automat legături între diverse concepte. În plus, unele elemente din pagini sunt semi-structurate și sunt ușor de prelucrat (de exemplu, infobox-urile). În continuarea acestei secțiuni, sunt prezentate câteva alternative de folosire a Wikipedia pentru a determina similaritatea semantică între diverse concepte.

Semantic Wikipedia

Semantic Wikipedia [4] este un proiect care încearcă să combine proprietățile Web-ului semantic cu cele ale paginilor Wiki. Pentru realizarea relațiilor semantice, se oferă utilizatorului posibilitatea de a adnota textul cu relații explicite între concepte. Un exemplu concret este cel din pagina de Wikipedia despre Londra [5]:

```
'''London''' is the capital city of
[[England]] and of the [[United Kingdom]].
```

În sintaxa Wikipedia, parantezele duble [[]] au semnificația de legătură către articolul cu titlul specificat între parantezele pătrate. Pentru afișare, parserul va transforma acest tip de legătură în hyperlink (în cod html). Pentru utilizator este clar faptul că Londra este capitala Angliei. Pentru o mașină însă, fiind limbaj natural, ar trebui aplicat un algoritm destul de complicat care realizează analiza sintactică a propoziției și inferează aceste rezultate. Semantic Wikipedia rezolvă problema aceasta altfel, prin adnotare textul de mai sus devenind:

```
'''London''' is the capital city of
[[capital of::England]] and of the
[[capital of::United Kingdom]].
```

Semnificația este aceea că 'London' are proprietatea 'capital of' cu valoarea 'England', respectiv 'United Kingdom'.

Semantic Wikipedia promite a fi o ontologie de mari dimensiuni, care va îmbunătăți semnificativ rezultatele în Web-ul semantic. Singurul dezavantaj este faptul că implică efortul uman pentru construirea relațiilor semantice.

Determinarea distanței semantice folosind Wikipedia

Pe aceasta temă au fost scrise mai multe lucrări, dar nu vor fi prezentate în această secțiune decât cele mai semnificative dintre acestea. De exemplu, Gabrilovich și Markovitch [6] propun ca pentru calculul distanței semantice dintre două concepte să se folosească vectorii spațiali ai articolelor corespunzătoare din Wikipedia, folosind distanța cosinus între aceștia. Metoda are rezultate destul de bune din punct de vedere al relevanței, problema principală este însă că Wikipedia are peste $N = 2$ milioane de articole și pentru a calcula pentru fiecare

articol vectorul *tf-idf*, iar apoi o complexitate de $O(N^2)$ pentru calculul distanței cosinus între toate conceptele este greu de implementat.

O altă încercare de a determina distanța semantică este WikiRelate [7]. În principiu, aceasta determină distanța dintre categoriile cărora aparțin conceptele. Ca și la abordarea precedentă, problema majoră este necesitatea efectuării unui calcul de 2 milioane x 2 milioane articole.

ÎMBUNĂȚIREA REZULTATELOR CĂUTĂRIILOR WEB FOLOSIND WIKIPEDIA

Pentru implementarea sistemului de reclasificare a rezultatelor întoarse de un motor de căutare web folosind legăturile semantice între conceptele din Wikipedia, s-a pornit de la modificarea modelului Google standard de clasificare. Acesta se bazează pe ordonarea paginii în funcție de următoarea formula [8]:

- Pentru interogările care sunt formate dintr-un singur cuvânt:

$$\text{Rank} = \text{PageRank} + \text{IR score} (\text{type-weight} (\text{content}, \text{title}, \text{etc}) + \text{term-weight})$$

- Pentru interogările formate din mai multe cuvinte cheie:

$$\text{Rank} = \text{PageRank} + \text{IR score} (\text{type-proximity} - \text{weight})$$

În momentul în care utilizatorul trimite o interogare către un server Google, un lexicon face asocierea dintre cuvintele și ID-uri. Având un index invers calculat în prealabil, se determină documentele în care apar cuvinte cu ID-urile determinate de lexicon. Aceste pagini sunt întoarse utilizatorului într-o ordine determinată de importanța paginii (determinată cu ajutorul algoritmului PageRank) precum și de un factor rezultat din mineritul datelor. Acest factor se obține luând în considerare numărul de apariții al cuvintelor cheie în pagină, locul în care apar acestea (cele din titlu de exemplu sunt considerate mai importante), fontul cu care sunt scrise și proximitatea dintre acestea (dacă interogarea este formată din mai multe cuvinte).

Se observă faptul că în modelul Google nu se ține cont de semnificația semantică a cuvintelor. Prin urmare sunt șanse ca o pagină cu mai puțină informație să fie afișată pe primele locuri doar pentru simplul fapt că este o pagină importantă (pagina unei companii cunoscute, de exemplu) și apar cuvintele cheie des în text.

Aplicația dezvoltată își propune calculul ordinii pentru afișarea rezultatelor folosind o pondere între scorul întors de motorul de căutare ($\text{Rank}(\text{Google})$) și scorul semantic, aplicând procedeul următor:

$$\text{Rank} = \text{Rank}(\text{Google}) + p * S \text{ score}$$

Unde $S \text{ score}$ este o valoare obținută pe baza informațiilor din Wikipedia, iar p este ponderea cu care această informație influențează scorul întors de motorul de căutare. Cu cât ponderea este mai mare, cu atât rezultatele semantice vor avea o influență mai mare.

Arhitectura propusă este prezentată în figura 1. Se remarcă două etape de funcționare: o etapă de preprocesare care este independentă de cuvintele cheie introduse de utilizator și o etapă de procesare a rezultatelor primite în

funcție de interogarea dată de utilizator și de relațiile semantice obținute în etapa anterioară.

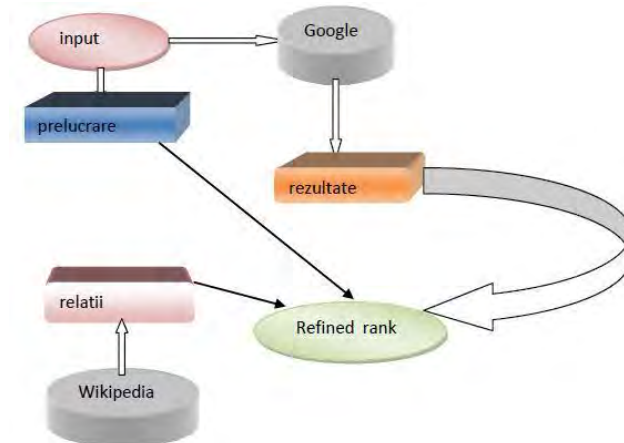


Figura 1. Arhitectura aplicației dezvoltate pentru reordonarea rezultatelor întoarse de Google folosind informații semantice

Etapă de preprocesare pentru calculul similarității semantice între conceptele din Wikipedia

Această etapă realizează prelucrări asupra articolelor din Wikipedia, încercând să obțină relații semantice pe baza legăturilor interne dintre articolele acestora precum și a paginilor de redirectare. Se consideră că fiecare articol diferit reprezintă un concept, iar legăturile web între articole exprimă o oarecare similaritate semantică între concepte, chiar dacă relația semantică nu este definită explicit. Acest lucru este posibil datorită structurii dense a enciclopediei, remarcându-se o multitudine de legături interne între articole.

Wikipedia poate fi văzută ca un graf, nodurile fiind articolele, iar arcele legăturile web între articole. Arcele au asociată o anumită pondere care reprezintă cât de puternică este legătura din punct de vedere semantic. Pentru determinarea relațiilor se folosește un algoritm similar cu *tf-idf* (*term frequency - inverse document frequency*), denumit *Path Frequency - Backward Link Frequency* [9]. Aceasta metodă a fost dezvoltată și optimizată pentru Wikipedia și calculează nu doar relațiile dintre vecini, ci și între noduri între care este distanța de n hopuri.

Pentru a reține graful se folosește o structură de date numită Arbore Binar Dual - *Dual Binary Tree* [9]. Această structură speculează faptul că matricea de adiacență este rară. În acest caz se poate folosi un arbore binar echilibrat (implementat ca un arbore AVL) pentru a reține liniile din matrice (acesta se numește *i-tree*). Fiecare intrare va avea un pointer spre rădăcina altui arbore AVL care conține efectiv elementele de pe linie (*j-tree*).

Pornind de la matricea de legături dintre articolele din Wikipedia, pentru calculul *pf-blf* se definește o nouă matrice în felul următor:

$$A' = A + A^T$$

Se observă că această matrice evidențiază atât legăturile directe ale unui articol cu vecinii săi, cât și legăturile inverse. Matricea va fi reținută folosind structura *dual binary tree* descrisă mai devreme. Pentru a determina relațiile între articole care se află la distanța de n hopuri,

este necesar sa se ridice această matrice la puterea n . În [9], este definit și un algoritm eficient pentru ridicarea la putere folosind structura *DBT*.

Algorithm MultiplyDBT(A)

```

1 for  $i \in i\text{-Tree}$ 
2 for  $j \in j\text{-Tree}(i)$ 
3 for  $k \in j\text{-Tree}(j)$ 
4  $R_{i,k} := R_{i,k} + a_{j,k} \cdot a_{i,j}$  ;
```

În algoritmul de mai sus, funcția $j\text{-Tree}(i)$ extrage toate elementele de pe linia i . Complexitatea algoritmului este, în principiu, $O(M^2N \log N)$, unde M este numărul de legături și N numărul de articole. Pentru calculul distanței $pf\text{-}blf$, înainte de a se ridica matricea la puterea n , se înlocuiește fiecare termen al acesteia cu formula:

$$a_{i,j}^n = a_{i,j} \cdot \log \frac{N}{|B_{vj}|}$$

unde $|B_{vj}|$ este numărul de muchii înapoi ale articolului cu indicele j . Folosind această matrice, precum și matricile calculate la puterea 2, 3, ..., n , se poate defini distanța finală între două articole/concepte din Wikipedia astfel:

$$pfibf(i, j) = \sum_{l=1}^n \frac{1}{d(n)} * a_{i,j}^l$$

unde $d(n)$ este o funcție crescătoare oarecare care scade influența conceptelor aflate la distanță mai mare de 1 în formula similarității semantice. Pentru a nu scade foarte repede aceste valori, se poate alege o funcție lent crescătoare, de exemplu, o funcție de tip logaritmic.

Astfel, în urma acestei etape de procesare se determină relațiile între articolele din Wikipedia. Într-o prima etapă de dezvoltare se poate considera că fiecare articol este reprezentat de conceptele din titlu. Ulterior, pot fi aplicate diverse metode pentru a determina pe lângă conceptele din titlu, și alte concepte importante din articol. Astfel de metode se pot baza pe frecvența apariției conceptelor în articol sau pe analiza părților de propoziție.

Astfel, în urma acestei preprocesări se va obține pentru fiecare concept important din Wikipedia, o listă de concepte cu care este relaționat precum și cat de puternică este relația între acestea ca un scor. Inițial, conceptele importante sunt cuvintele din titlu pe care se aplică un algoritm de stemming [10]. Ținând cont de structura Wikipedia și de modelarea datelor (legăturile dintre articole sunt stocate într-o tabelă separată în baza de date) distanța $pf\text{-}idf$ se dovedește a fi potrivită pentru extragerea de relații semantice.

Etapa de prelucrare a textului interogării

În momentul în care utilizatorul introduce cuvintele cheie pentru căutare, acestea sunt trimise la Google și preluate folosind Google API. Din interogarea precizată de utilizator și din rezultatele întoarse de Google se elimină apoi cuvintele de stop [11] și se aplică un algoritm de stemming. Pentru fiecare cuvânt cheie, se caută atât cuvântul respectiv, cât și conceptele cu care acesta este în relație mai strânsă în fiecare document web întors de motorul de căutare. Pentru eficiență, se creează în prealabil un index care conține pentru fiecare document, cuvintele din interiorul său și numărul de apariții ale

acestora. Se obține un nou scor în funcție de importanța conceptului (cuvintele cheie din interogare au importanță maximă) în funcție de formula următoare:

$$Score_{k,vj} = \sum tf_{i,vj} * tfblf_{k,i}$$

unde k este un cuvânt cheie din interogare, vj este un document web întors de motorul de căutare, iar i este un concept cu care cuvântul cheie este în relație (se consideră că fiecare cuvânt cheie este într-o relație foarte strânsă cu el însuși).

Pornind de la relația anterioară, scorul semantic final pentru un document ține cont de similaritatea semantică între document și toate cuvintele cheie din interogare, folosind formula peste $Score$:

$$Score_{vj} = \sum_k Score_{k,vj}$$

Ordinea finală va fi PageRank-ul determinat de Google la care se adaugă cu o anumită pondere scorul semantic obținut. Ponderile folosite în formulele prezentate sunt determinate empiric. Totuși, pentru rezultate cât mai bune, este nevoie de o testare în aprofundime, eventual pe câmpuri semantice diferite pentru a obține rezultate satisfăcătoare pentru setarea acestei ponderi.

SCENARIU DE UTILIZARE

Pornim de la următorul exemplu concret: utilizatorul introduce interogarea „black tree” și dorește o căutare semantică în domeniul ingineriei calculatoarelor. În această situație, se determină două câmpuri semantice diferite: pot fi considerate relevante paginile despre structurile de date arbori roșu-negru des întâlnite în implementarea multor algoritmi și, eventual, paginile despre arbori ca plante. Doar primul câmp semantic este relevant pentru utilizatorul nostru. Cum însă nu există o specie de copaci „arbori negri” sau cel puțin nu există una cunoscută la scară largă, se așteaptă ca primele rezultate întoarse de Google să fie despre structura de date menționată. Într-adevăr, conceptul „black tree” nu este unul „pur”, interogarea completă fiind „red black tree”. Însă, pentru conceptul „black tree” cel mai apropiat concept este „red black tree”, explicându-se astfel așteptările utilizatorului. Un potențial utilizator este de exemplu un student în domeniul calculatoarelor care nu a reținut denumirea completă a structurii de date și dorește să obțină mai multe informații despre aceasta.

Totuși, primul rezultat întors de motorul de căutare este către pagina web: <http://www.blacktree.com/>. Aceasta pagină conține proiecte OpenSource pentru sistemele Mac. Conținutul semantic este foarte redus, indiferent de domeniul considerat, fiind folosit doar scorul lexical unde este luat în considerare titlul său corespunde cuvintelor din interogarea introdusă de utilizator. Similar, în primele 10 pagini apar <http://www.black-tree-design.com/> și <http://www.blacktreedesign.com>, care se referă la firme cu același nume, dar care au domenii web și de activitate diferite. De pe aceste două pagini web, utilizatorul are posibilitatea să cumpere statuiete și diverse obiecte de decor. Se poate remarca faptul că cele două pagini au informații redundante. Totodată, aceste pagini sunt utile în

cazul unei căutări tranzacționale, dar nu și pentru căutări informaționale.

Al cincilea rezultat întors de Google este pagina: http://en.wikipedia.org/wiki/Red-black_tree, care corespunde articolului despre arbori roșu și negru din Wikipedia. Pagina conține o mulțime de informații despre acești arbori precum și despre concepte înrudite cu acesta: arbori binari de căutare, arbori binari echilibrați. În text apar de multe ori cuvintele din interogarea utilizatorului, precum și cuvinte din câmpul semantic înrudit denumirii acestei structuri de date.

Pentru scenariul nostru de utilizare, se consideră că pagina aceasta este mai relevantă din punct de vedere semantic pentru utilizatorul interesat de ingineria calculatoarelor care a introdus interogarea: *black tree*. Se dorește ca motorul de căutare să ofere această pagină utilizatorului în primele locuri, rezultate precum cele menționate mai sus despre firma de design Black Tree fiind mai puțin relevante. Pentru a determina dacă un document conține informații despre un concept introdus de utilizator este necesară prelucrarea textului luând în considerare și concepte relaționate cu cuvintele cheie introduse de utilizator.

Vom prezenta în continuare mai multe rezultate obținute în cadrul acestui scenariu de utilizare.

REZULTATE OBȚINUTE FOLOSIND REORDONAREA PAGINILOR ÎNTOARSE DE MOTORUL DE CĂUTARE

Pentru testarea abordării propuse în această lucrare, s-a pornit de la scenariul de utilizare prezentat anterior. Pentru că există un număr foarte mare de pagini și legături între ele în Wikipedia, s-au calculat legăturile semantice între toate articolele din enciclopedie în etapa de preprocesare, dar sunt extrase și folosite numai conceptele legate cu domeniul „*computer*” sau concepte care conțin acest cuvânt.

Înainte de a prezenta rezultatele finale obținute după aplicarea reordonării paginilor web, este interesant de prezentat rezultatele etapei de preprocesare. În tabele 1 și 2 sunt prezentate cele mai apropiate concepte referitoare la conceptele sursă „*avl tree*” și „*computer hardware*”. Astfel, se observă că conceptele destinație cu scorul cel mai mare obținut în urma aplicării algoritmului *pf-lbf* sunt într-adevăr concepte apropiate semantic de către conceptele sursă și care pot influența în mod pozitiv căutarea.

Tabelul 1. Relații semantice pentru conceptul sursă „*avl tree*” împreună cu ponderile calculate

Concept sursă	Concept destinație	Pondere
avl tree	self balanced binary tree	0.703385
avl tree	amortization analysis	0.605471
avl tree	donald knuth	0.557814
avl tree	b tree	0.546726
avl tree	binary tree	0.527878
avl tree	computer science	0.527694
avl tree	persistent data structure	0.50078
avl tree	red black tree	0.441281
avl tree	associative array	0.437683

Tabelul 2. Relații semantice pentru conceptul sursă „*computer hardware*” împreună cu ponderile calculate

Concept sursă	Concept destinație	Pondere
computer hardware	computer network	0.264669
computer hardware	dell	0.247488
computer hardware	hewlett packard	0.237757
computer hardware	red hat	0.231325
computer hardware	lg electronics	0.228782
computer hardware	general electric	0.227617
computer hardware	apple inc.	0.227089
computer hardware	novel	0.226752
computer hardware	computer storage	0.22162

Afișarea rezultatelor întoarse de aplicație se face într-o pagină web cu o interfață similară cu aceea a Google, folosind o paginare în care sunt prezentate câte 8 snippeturi către site-urile web considerate conform reordonării ce folosește informația semantică. Trebuie avut în vedere că pentru reclasificare sunt considerate doar primele 60 de rezultate întoarse de către motorul de căutare. Se consideră astfel că în primele 60 de rezultate trebuie să fie prezente și pagini care au multe concepte din aria semantică de interes pentru utilizator. Oricum, sunt șanse foarte mici ca paginile aflate mai departe de locul 60 în ordonarea oferită de Google să urce în primele locuri după considerarea informației semantice.

Chiar și în această situație, trebuie avut în vedere că timpul de procesare este destul de ridicat pentru a fi rulate pe un calculator obișnuit: Core2 Duo 2.0GHz, 2 GB RAM, viteza medie de download 120 KB/sec. În continuare, pentru anumite interogări introduse de utilizator este afișată durata procesării acestora:

Computer engineer - 24.30 sec
Black Tree - 34.39 sec
Dijkstra - 31.54 sec
Kruskal - 25.47 sec

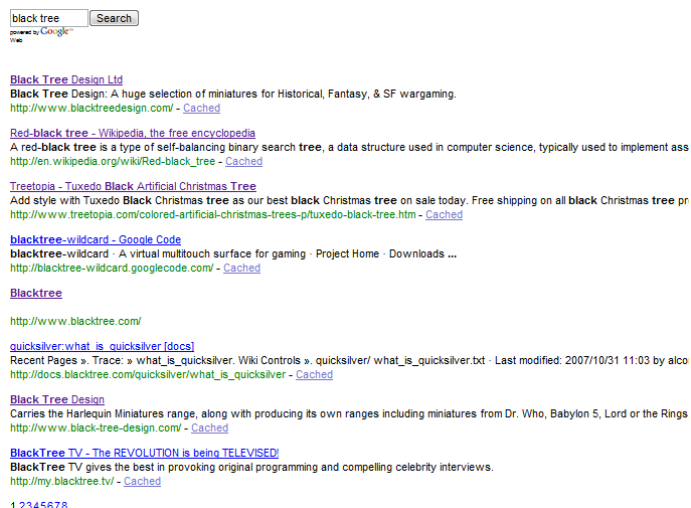


Figura 2. Rezultatele întoarse de către aplicație pentru interogarea „*black tree*”, după aplicarea reordonării paginilor ținând cont de scorul semantic

Evaluarea rezultatelor obținute este dificil de efectuat, cel mai relevant test fiind unul de satisfacere a nevoilor utilizatorilor. În figurile 2 și 3 sunt prezentate rezultatele comparative pentru căutarea prezentată în scenariul de utilizare: *black tree*. Comparativ cu rezultatele oferite de Google, se poate observa faptul că pagina Wikipedia a

crescut în rang fiind afișată pe locul 2, așa cum este de dorit în cadrul scenariului de utilizare. Din păcate, primul rezultat este firma de design care are un scor destul de bun datorită PageRank-ului ridicat și apariției cuvintelor cheie într-un context în care sunt prezente puține cuvinte. Interesant este rezultatul de pe locul 3, despre o firmă care oferă brazi de Crăciun artificiali negri (în engleză, *black Christmas tree*), rezultat perfect valabil din punct de vedere semantic (*tree* fiind puternic legat de *Christmas tree*).

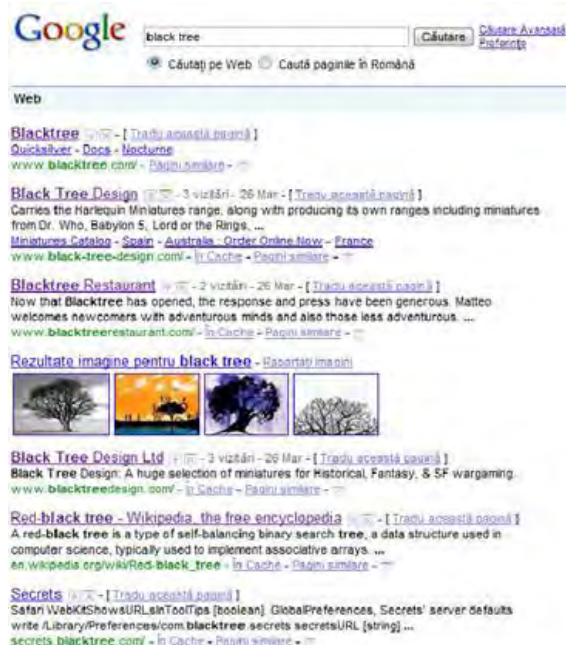


Figura 3. Rezultatele întoarse de Google pentru interogarea „black tree”

CONCLUZII

Wikipedia este o sursă extrem de valoroasă pentru extragerea informațiilor semantice, însă sunt întâmpinate dificultăți datorate în special dimensiunii foarte mari și a conținutului semistructurat pentru a obține legături semantice explicite cu precizie foarte bună. Lucrarea de față folosește un algoritm pentru extragerea legăturilor semantice între articolele din Wikipedia pentru a ajuta utilizatorul să ajungă cât mai rapid la informația căutată pe web. Pentru aceasta, se combină scorul întors de motorul de căutare cu scorul semantic calculat folosind Wikipedia și se reordonează rezultatele folosind acest scor ponderat.

Pentru validare s-a considerat un profil al utilizatorului interesat de ingineria calculatoarelor, folosind relațiile semantice găsite între articolele din Wikipedia legate de

acest domeniu. Rezultatele testării sistemului sunt îmbucurătoare, paginile reclasificate conform metodei propuse fiind mai relevante pentru utilizator.

REFERINȚE

1. Manning, C. D, Raghavan, P.,Schutze, H.: Introduction to Information Retrieval. (eds). Cambridge University Press (2008)
2. Hao, T., Lu, Z., Wang, S., Zou, T., Gu, S., Wenyin, L.: Categorizing and ranking search engine's results by semantic similarity, at Conference On Ubiquitous Information Management And Communication (2008)
3. Haveliwala, T.H.: Topic-Sensitive PageRank: A context-sensitive Ranking Algorithm for Web Search. 11th International World Wide Web Conference (2002)
4. Semantic MediaWiki - http://semantic-mediawiki.org/wiki/Semantic_MediaWiki
5. Kotzsh, M., Vrandecic, D., Volkel, M., Haller H., Studer, R.: Semantic Wikipedia (2007)
6. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. În: Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2007), 1606–1611(2007)
7. Strube, M., Ponzetto, S.: WikiRelate! Computing semantic relatedness using Wikipedia. În: Proc. Of National Conference on Artificial Intelligence (AAAI-06), Boston, Mass. (2006) 1419–1424
8. The Google Project at Stanford
<http://infolab.stanford.edu/~backrub/google.html>
9. Nakayama, K., Hara, T., Nishio, S.: Wikipedia mining for an association web thesaurus construction, în Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007), pp. 322–334 (2007)
10. <http://en.wikipedia.org/wiki/Stemming>
11. http://www.dcs.gla.ac.uk/ir_resources/linguistic_utils/stop_words
12. Müller, C., Gurevych, I.: Using Wikipedia and Wiktionary în Domain - Specific Information Retrieval. În: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark, September 17 - 19, 2008 (2008)
13. Border, A., Ciccolo, P., Gabrilovich, E.: Online Expansion of Rare Queries for Sponsored Search
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab (1999)