Îmbunătăţirea relevanţei rezultatelor motoarelor de căutare folosind informaţii semantice din Wikipedia

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

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

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

Ştefan Trăuşan-Matu
Universitatea „Politehnica” din Bucureşti,
Splaiul Independenței, Nr 313 Bucureşti
Institutul de Cercetare în Inteligenţa Artificială al Academiei Române,
Calea 13 Septembrie, Nr 13 Bucureşti
trausan@cs.pub.ro

REZUMAT

În funcţie de intenţia utilizatorului, interogările către un motor de căutare pot fi împărţite în trei categorii: tranzacţionale, informaţionale şi navigaţionale [1]. Pentru a satisface cele trei tipuri de căutări, motoarele de căutare din prezent folosesc în principiu algoritmi de analiză a legăturilor între pagini îmbunătăţiţi cu un factor care depinde de numărul de apariţii şi ordinea cuvintelor căutate. Pentru interogările tranzacţionale şi informaţionale, relevanţa ar putea fi îmbunătăţită dacă s-ar folosi şi informaţii 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 informaţii în diverse moduri. Lucrarea de faţa propune o modalitate de extragere a relaţiilor semantice din Wikipedia, iar apoi folosirea acestora pentru a determina ordinea în care rezultatele întoarse de un motor de căutare sunt afişate utilizatorului.

Cuvinte cheie

Căutare web, motor de căutare, optimizare, relaţii 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 şi diferite idei au fost sugerate, atât din mediul academic cât şi din industrie. În prezent, utilizatorii pot să obţină diferite rezultate în funcţie de cuvintele cheie pe care le furnizează motorului de căutare. Rezultatele obţinute sunt destul de relevante, însă se observă o lipsă a relaţiilor semantice, ordonarea făcându-se după importanţa paginii şi conţinutul lexical şi nu după conţinutul semantic. Mai mult, sunt multe informaţii redundante în aceste pagini, ceea ce măreşte timpul de căutare al utilizatorului. În aceste condiţii, această lucrare propune o reordonare a rezultatelor aduse de un motor de căutare (de exemplu, Google), în funcţie de conţinutul semantic.

Pentru extragerea de informaţii semantice pot fi folosite două tipuri de baze de cunoştinţe: baze de cunoştinţe lingvistice (dicţionare / ontologii precum WordNet) sau baze de cunoştinţe colaborative precum Wikipedia. Folosind baze de cunoştinţe lingvistice, au fost numeroase încercări de expansiune a cuvintelor cheie date de utilizator pentru a obţine 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, restricţionarea la un vocabular ce conţine concepte generale în mare parte.

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

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

Metoda propusă pentru determinarea relaţiilor semantice este următoarea: Pe baza legăturilor înainte şi înapoi dintre articole, se determină distanţa dintre articole. Pentru început, se consideră că articolul reprezintă conceptul din titlu, obţinându-se astfel o relaţie între concepte. În momentul în care utilizatorul introduce cuvintele cheie, se foloseşte un motor de căutare pentru a obţine 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 şi cuvintele aflate în relaţie cu cuvintele

cheie prelucrate. În funcţie de numărul de apariţii ale acestor concepte precum şi de importanţa lor se obţine un nou scor care este combinat cu cel determinat de Google. Cuvintele cheie introduse de utilizator au importanţă maximă, iar restul conceptelor au o importanţă determinată în funcţie de cât de puternică este legătura dintre acestea şi un cuvânt cheie introdus de utilizator, folosind Wikipedia.

Restul lucrării este structurat în modul următor: în secţiunea următoare sunt prezentate alte metode de îmbunătăţire a relevanţei rezultatelor motoarelor de căutare precum şi alte sisteme de extragere a informaţiilor semantice folosind Wikipedia. În secţiunea 3 este descris în detaliu modelul propus, iar în secţiunea 4 sunt prezentate rezultatele obţinute. În final, vor fi prezentate concluziile.

ABORDĂRI ANTERIOARE PENTRU ÎMBUNĂTĂŢIREA REZULTATELOR CĂUTĂRILOR WEB FOLOSIND INFORMAŢII SEMANTICE

Recalcularea ordinii rezultatelor unui motor de căutare web folosind WordNet

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

Algoritmul de calcul al similarităţii semantice ales se bazează pe WordNet, calculând în principiu lungimea căii celei mai apropiate dintre cele două concepte pentru care se doreşte determinarea distanţei semantice (ţinând cont de hiponime) precum şi cantitatea de informaţie conţinută de concept, care depinde de probabilitatea de apariţie a conceptului în întreg tezaurul.

În a doua parte a lucrării se propune un nou algoritm pentru ordonarea rezultatelor. Astfel, noul scor se obţine prin combinarea PageRank-ului [14] obţinut de Google cu factorul de similaritate dintre snippets şi topicul din interogare precum şi cu un factor de timp. Pentru calculul factorului de timp se porneşte de la premisa că utilizatorul doreşte sa vadă cea mai nouă informaţie. Astfel factorul este invers proporţional cu diferenţa dintre timpul curent şi timpul de cache.

Algoritmul Topic-Sensitive PageRank

O altă abordare porneşte de la ideea că este mai bine să se folosească un algoritm similar cu PageRank-ul clasic. În încercarea de a îmbunătăţi algoritmul PageRank clasic cu informaţii 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 mulţi vectori, fiecare pentru un topic anume. Utilizatorului îi sunt afişate rezultatele în funcţie de vectorul topicului interogării introdus. Topicul interogării poate fi determinat fie din context, fie prin calculul distanţei faţa de toţi vectorii, alegându-se cel cu distanţa cea mai mica.

Utilizarea Wikipedia ca sursă de informaţii semantice

În ultima perioadă, ideea utilizării Wikipedia ca sursa de informaţii semantice este din ce in ce mai vehiculată. În primul rând, Wikipedia conţine informaţii relevante pentru majoritatea domeniilor şi conceptelor prezentate. În al doilea rând, numărul de concepte şi domenii acoperite este foarte mare, depăşind din punct de vedere al numărului de pagini orice alternativă. Poate că cel mai important avantaj este faptul că se foloseşte 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 şi sunt uşor de prelucrat (de exemplu, infobox-urile). În continuarea acestei secţiuni, 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 proprietăţile Web-ului semantic cu cele ale paginilor Wiki. Pentru realizarea relaţiilor semantice, se oferă utilizatorului posibilitatea de a adnota textul cu relaţii 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 semnificaţia de legătură către articolul cu titlul specificat între parantezele pătrate. Pentru afişare, parserul va transforma acest tip de legătură în hyperlink (în cod html). Pentru utilizator este clar faptul că Londra este capitala Angliei. Pentru o maşina însă, fiind limbaj natural, ar trebui aplicat un algoritm destul de complicat care realizează analiza sintactică a propoziţiei şi 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]].
```

Semnificaţia 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 îmbunătăţi semnificativ rezultatele în Web-ul semantic. Singurul dezavantaj este faptul că implică efortul uman pentru construirea relaţiilor semantice.

Determinarea distanţei semantice folosind Wikipedia

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

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

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

ÎMBUNĂTĂȚIREA REZULTATELOR CĂUTĂRILOR 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 funcţie de următoarea formula [8]:

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

```
Rank = PageRank + IR score (type-weight (content, title, etc) + term-weight)
```

- Pentru interogările formate din mai multe cuvinte cheie:

```
Rank = PageRank + IR score ( type-proximity – weight)
```

În momentul în care utilizatorul trimite o interogare către un server Google, un lexicon face asocierea dintre cuvintele şi 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 importanţa paginii (determinată cu ajutorul algoritmului PageRank) precum şi de un factor rezultat din mineritul datelor. Acest factor se obţine luând în considerare numărul de apariţii al cuvintelor cheie în pagină, locul în care apar acestea (cele din titlu de exemplu sunt considerate mai importante) , fontul cu care sunt scrise şi proximitatea dintre acestea (dacă interogarea este formată din mai multe cuvinte).

Se observă faptul că în modelul Google nu se ţine cont de semnificaţia semantică a cuvintelor. Prin urmare sunt şanse ca o pagină cu mai puţină informaţie să fie afişată pe primele locuri doar pentru simplul fapt că este o pagină importantă (pagina unei companii cunoscute, de exemplu) şi apar cuvintele cheie des in text.

Aplicaţia dezvoltată îşi propune calculul ordinii pentru afişarea rezultatelor folosind o pondere între scorul întors de motorul de căutare (*Rank(Google)*) şi scorul semantic, aplicând procedeul următor:

```
Rank = Rank(Google) + p * S score
```

Unde *S score* este o valoare obţinută pe baza informaţiilor din Wikipedia, iar *p* este ponderea cu care această informaţie influenţează scorul întors de motorul de căutare. Cu cât ponderea este mai mare, cu atât rezultatele semantice vor avea o influenţă mai mare.

Arhitectura propusă este prezentată în figura 1. Se remarcă două etape de funcţionare: o etapă de preprocesare care este independentă de cuvintele cheie introduse de utilizator şi o etapă de procesare a rezultatelor primite în funcţie de interogarea dată de utilizator şi de relaţiile semantice obţinute în etapa anterioară.

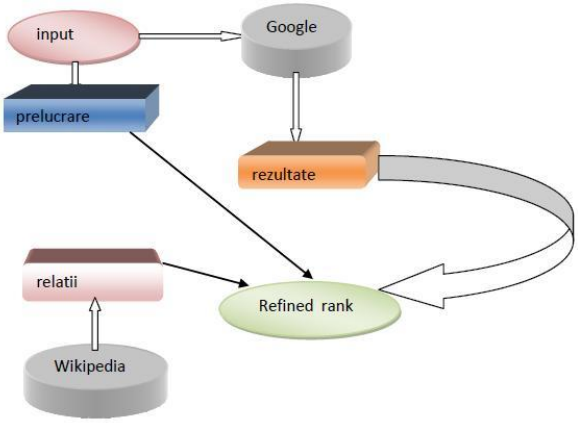

Figura 1. Arhitectura aplicației dezvoltate pentru reordonarea rezultatelor întoarse de Google folosind informații semantice

Etapa de preprocesare pentru calculul similarității semantice între conceptele din Wikipedia

Aceasta etapă realizează prelucrări asupra articolelor din Wikipedia, încercând să obţină relaţii semantice pe baza legăturilor interne dintre articolele acesteia precum şi 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 relaţia 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 relaţiilor se foloseşte un algoritm similar cu *tf-idf* (*term frequency - inverse document frequency*), denumit *Path Frequency - Backward Link Frequency* [9]. Aceasta metodă a fost dezvoltată şi optimizată pentru Wikipedia şi calculează nu doar relaţiile dintre vecini, ci şi între noduri între care este distanţa de n hopuri.

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

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

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

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

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

```
Algorithm MultiplyDBT(A)
  1 for i ∈ i-Tree
    2 for j ∈ j-Tree(i)
      3 for k ∈ j-Tree(j)
        4 Ri,k := Ri,k + aj,k · ai,j ;
```

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

$$a'_{i,j} = a'_{i,j} \cdot \log\frac{N}{|Bvj|}$$

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

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

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 logaritm.

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, şi 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 şi 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]. Ţinând cont de structura Wikipedia şi de modelarea datelor (legăturile dintre articole sunt stocate într-o tabelă separată in baza de date) distanța *pf-idf* se dovedeşte 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 şi preluate folosind Google API. Din interogarea precizată de utilizator şi din rezultatele întoarse de Google se elimină apoi cuvintele de stop [11] şi se aplică un algoritm de stemming. Pentru fiecare cuvânt cheie, se caută atât cuvântul respectiv, cât şi 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 şi 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:

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

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 însuşi).

Pornind de la relația anterioară, scorul semantic final pentru un document ține cont de similaritatea semantică între document şi toate cuvintele cheie din interogare, folosind formula peste *S score*:

$$S\ score_{vj} = \sum_k S\ score_{k,vj}$$

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. Totuşi, pentru rezultate cât mai bune, este nevoie de o testare in aprofunzime, 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*" şi doreşte 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 roşu-negru des întâlnite în implementarea multor algoritmi şi, 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 aşteaptă 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 aşteptările utilizatorului. Un potențial utilizator este de exemplu un student în domeniul calculatoarelor care nu a reținut denumirea completă a structurii de date şi doreşte sa obțină mai multe informații despre aceasta.

Totuşi, 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/ şi http://www.blacktreedesign.com, care se referă la firme cu acelaşi nume, dar care au domenii web şi de activitate diferite. De pe aceste două pagini web, utilizatorul are posibilitatea să cumpere statuiete şi diverse obiecte de decor. Se poate remarca faptul că cele două pagini au informații redundante. Totodată, aceste pagini sunt utile in

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 aceşti arbori precum şi 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 1și 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	hewllet 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 sa 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
```

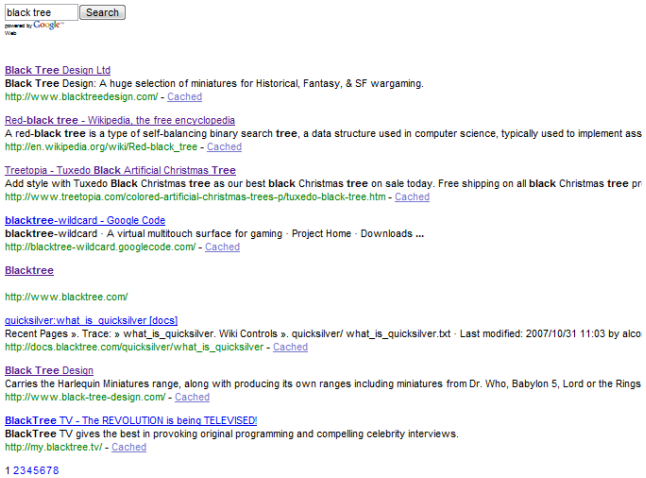

Figura 2. Rezultatele întoarse de către aplicație pentru interogarea „black tree", după aplicarea reordonării paginilor ținând cont de scorul semantic

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 afişată pe locul 2, aşa 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 şi apariţiei cuvintelor cheie într-un context în care sunt prezente puţine 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*).

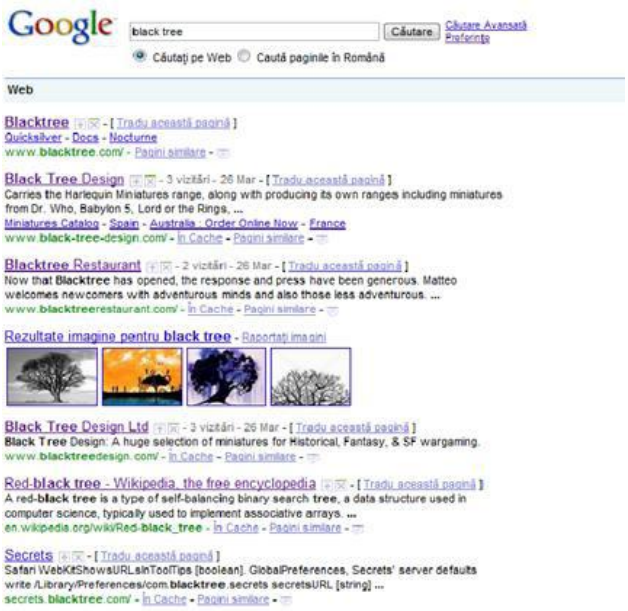

Figura 3. Rezultatele întoarse de Google pentru interogarea „black tree”

CONCLUZII

Wikipedia este o sursă extrem de valoroasă pentru extragerea informaţiilor semantice, însă sunt întâmpinate dificultăţi datorate în special dimensiunii foarte mari şi a conţinutului semistructurat pentru a obţine legături semantice explicite cu precizie foarte bună. Lucrarea de faţa foloseşte un algoritm pentru extragerea legăturilor semantice între articolele din Wikipedia pentru a ajuta utilizatorul sa ajungă cât mai rapid la informaţia căutată pe web. Pentru aceasta, se combină scorul întors de motorul de căutare cu scorul semantic calculat folosind Wikipedia şi se reordonează rezultatele folosind acest scor ponderat.

Pentru validare s-a considerat un profil al utilizatorului interesat de ingineria calculatoarelor, folosind relaţiile 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.

REFERINŢE

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/idom/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.: Qnline 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)