Concepte teoretice utilizate în regasirea textelor dupa cuvinte indexate si dupa tematici de cautare

Cerc.st. Anton-Mugurel ROG
Institutul pentru Tehnologii Avansate, Bucuresti
e-mail: ranton@icon.roknet.ro

Articolul îsi propune prezentarea câtorva concepte care stau la baza scrierii motoarelor de cautare. Se abordeaza doua modalitati de regasire: dupa cuvinte indexate (cunoscute ca si cuvinte cheie) si dupa tematici (cunoscute si ca sintagme). Cele doua tipuri de regasire sunt orientate spre baze de date relationale sau orientate catre text, locale sau distribuite.
Cuvinte cheie: *regasire, motor, cautare, tematici, indexare, algoritm, concepte, sintagme.*

Introducere

Explozia bazelor de date orientate catre text, precum si necesitatea asigurarii unui acces rapid la informatiile pe care le contin, au condus la crearea unei multitudini de motoare de cautare (de exemplu, yahoo, amazon, kappa.pcnet.ro, cauta.ro etc.). Asa cum se stie, textul nu este considerat o informatie structurata, deci nu se pot utiliza algoritmi standard de cautare. De asemenea, nu se pot utiliza comenzi din limbajele de interogare a bazelor de date pentru regasirea rapida a informatiilor dorite. În cazul unei baze de date relationale, cautarea textelor care contin un anumit cuvânt cu ajutorul comenzii SELECT, se efectueaza secvential, astfel timpul de raspuns devenind foarte mare.

Regasire dupa cuvinte indexate

Se abordeaza în primul rând unele concepte teoretice specifice regasirii textelor dupa cuvinte indexate. Printre acestea se regasesc:

a) *Lista cuvintelor care vor fi indexate.* În functie de interesul beneficiarului si de domeniul stabilit se propune o lista de cuvinte spre indexare. De exemplu, pentru domeniul economic, trebuie introduse în lista si cuvintele: **economie, servicii, industrie, publicitate, produse, marfa** etc. Referirile ulterioare la aceasta lista se vor face prin utilizarea termenului INDEXLIST.

b) *Lista familiilor de cuvinte pentru cuvintele propuse spre indexare*. Exista posibilitatea ca la formularea unei cereri de regasire sa nu se utilizeze exact derivarea cuvântului care apare în text, cu toate acestea raspunsul la cerere trebuie sa întoarca toate textele ce contin cel putin un membru derivat din familia cuvântului cautat. Membrii familiei unui substantiv de gen feminin sunt: **sfsn** (substantiv feminin singular nearticulat), **sfsa** (substantiv feminin singular articulat), **sfsd** (substantiv feminin singular dativ), **sfpn** (substantiv feminin plural nearticulat), **sfpa** (substantiv feminin plural articulat), **sfpd** (substantiv feminin plural dativ). Aceleasi reguli se aplica si pentru genurile masculin si neutru. În cazul cuvântului *economie* avem derivarile **sfsn – economie**, **sfsa – economia**, **sfsd – economiei**, **sfpa – economii**, **sfpa – economiile**, **sfpd – economiilor**. Referirile ulterioare la aceasta lista se vor face prin utilizarea termenului FAMLIST.

c) *Lista cuvintelor ce vor fi omise de la indexare* contine prepozitii, conjunctii, interjectii, cuvinte cu frecventa mare de aparitie ce nu sunt interesante pentru indexare. Printre cuvintele omise se pot regasi: **a, si, la, le, cum, ce, da, nu, acea, acel, ai, al, am, ca, ci, de, ei, ne, sa, va, sunt, prea, urmatoarele** etc. Referirile ulterioare la aceasta lista se vor face prin utilizarea termenului STOPLIST.

d) *Algoritmul de indexare*. Rolul acestuia este de a codifica legaturile dintre cuvintele propuse spre indexare (INDEXLIST) si textele în care acestea apar, precum si de a ignora de la indexare cuvintele din STOPLIST. Se pleaca de la premisa ca

fiecare text are asociat un identificator unic, o secventa bine stabilita. Prin codificarea legaturii dintre cuvânt si textul în care apare, se întelege pastrarea asocierii dintre identificatorul unic al textului si cuvântul respectiv. Algoritmul are doua parti, una pentru crearea indexului initial (construit din textele existente la acel moment în baza de date), iar a doua pentru actualizarea ulterioara a acestuia (ori de câte ori se modifica cel putin un text). De asemenea, algoritmul contine procedurile de regasire (decodificare a asocierilor dintre cuvinte si identificatorii unici) a textelor dupa cuvintele indexate. O mare categorie de algoritmi de indexare este aceea unde primeaza viteza de raspuns în pofida spatiului ocupat de index. Domeniile în care se utilizeaza fiind INTRANET si INTERNET, si anume la scrierea portalurilor. O alta mare categorie de algoritmi au ca prioritate crearea unui index care ocupa cât mai putin spatiu si raspunde la cereri într-un tip rezonabil dar mai mare decât timpul specific primei categorii de algoritmi. Ultima categorie se utilizeaza în companiile IT de marime mijlocie si mica. Pentru a fi considerat un algoritm fiabil, timpul de raspuns la o cerere de regasire nu trebuie sa fie mai mare de 20 secunde, iar spatiul ocupat de indexul construit trebuie sa reprezinte aproximativ 30 – 40% din spatiul ocupat de texte. Daca totusi timpul necesar efectuarii unei regasiri este mai mare de 20 secunde, atunci raspunsul trebuie sa fie segmentat în mai multe parti, astfel încât prima parte sa fie disponibila în mai putin de 20 secunde. Pe un esantion de 100.000 texte crearea indexului initial de regasire trebuie sa dureze cel mult 3 ore. (Durata poate varia în functie de puterea server-ului si de sistemul de gestiune a bazelor de date.)

e) *Operatorii* utilizati la scrierea cererilor de regasire:

- AND – operator binar; raspunsul la o cerere **cuvânt1 AND cuvânt2** trebuie sa întoarca toate textele ce contin atât **cuvânt1** cât si **cuvântul2** indiferent de ordinea de aparitie a acestora în text.
- OR – operator binar; raspunsul la o cerere **cuvânt1 OR cuvânt2** trebuie sa întoarca toate textele ce contin cel putin unul dintre **cuvânt1** si **cuvântul2** indiferent de ordinea de aparitie a acestora în text.
- NOT – operator unar; raspunsul la o cerere **NOT cuvânt** trebuie sa întoarca toate textele ce nu contin **cuvânt**.
- () – parantezele au rolul de a schimba ordinea de efectuare a operatorilor.

Operatorii au urmatoarea ordine de prioritate: NOT, AND, OR.

Evaluarea oricarei cereri de regasire trebuie sa se poata efectua utilizând algoritmul POLONEZEI INVERSE. De exemplu, în cererea **(cuvânt1 OR cuvânt2) AND cuvânt3** se evalueaza mai întâi cererea **cuvânt1 OR cuvânt2** conform regulii expuse la operatorul **OR**, urmând ca multimea textelor regasite sa se intersecteze cu multimea textelor ce contin **cuvânt3.**

Orice cerere se poate construi utilizând urmatoarea gramatica regulata:

(1) E → **cuvânt**
(2) E → E **AND** E
(3) E → E **OR** E
(4) E → **NOT** E
(5) E → **(**E**)**,

unde E este o cerere de regasire.

O cerere concreta, care în limbaj natural ar fi: "Care sunt toate textele reclamelor despre fabricile de tigarete fara filtru cu sediul în Bucuresti ?" se traduce în limbajul operatorilor astfel: (FABRICILE **AND** (TIGARETE **OR** TUTUN)) **AND NOT** FILTRU **AND** BUCURESTI. Pentru demonstrarea corectitudinii gramaticii se va reconstrui cererea de mai sus prin aplicarea repetata a productiilor.

E → E **AND** E (2)
E → E **AND** E **AND** E (2)
E → E **AND NOT** E **AND** E (4)
E → (E) **AND NOT** E **AND** E (5)

E → (E **AND** E) **AND NOT** E **AND** E (2)
E → (E **AND** (E)) **AND NOT** E **AND** E (5)
E → (E **AND** (E **OR** E)) **AND NOT** E **AND** E (3)
E → (E **AND** (E **OR** E)) **AND NOT** E **AND** BUCURESTI (1)
E → (E **AND** (E **OR** E)) **AND NOT** FILTRU **AND** BUCURESTI (1)
E → (E **AND** (E **OR** TUTUN)) **AND NOT** FILTRU **AND** BUCURESTI (1)
E → (E **AND** (TIGARETE **OR** TUTUN))
AND NOT FILTRU **AND** BUCURESTI (1)
E → (FABRICILE **AND** (TIGARETE **OR** TUTUN))
AND NOT FILTRU **AND** BUCURESTI (1)

f) *Reguli de actualizare*. Orice actualizare a textelor, ulterioara crearii indexului initial, trebuie sa se reflecte automat (on-line) în acesta. Concret, prin actualizare se întelege:
STERGERE TEXT – atunci când se sterge un text, se sterg toate legaturile dintre cuvintele propuse spre indexare, ce se regasesc în acesta, si textul ca atare.
INSERARE TEXT – la inserarea unui text, se adauga toate legaturile dintre cuvintele propuse spre indexare, ce se regasesc în acesta, si textul ca atare.
MODIFICARE TEXT – modificarea unui text presupune aplicarea regulii de la stergere pentru forma initiala a textului si apoi a regulii de la inserare pentru forma de dupa modificare a textului.
Se recomanda ca modulele software care realizeaza actualizarea indexului sa ruleze sub forma unor procese în background astfel încât modificarile sa se produca automat si într-un timp cât mai scurt, asemanator trigger-ilor de la nivelul bazei de date.

Regasire dupa tematici

A doua parte a materialului face referire la câteva concepte teoretice specifice regasirii textelor dupa tematici. Dintre acestea se vor aborda:
a) *Lista cuvintelor ce vor compune sintagmele*. În functie de interesul beneficiarului se propune o lista de cuvinte din care se vor putea construi sintagmele specifice tematicilor de cautare. Referirile ulterioare la aceasta lista se vor face prin utilizarea termenului WORDLIST.
b) *Lista familiilor de cuvinte pentru cuvintele ce vor compune sintagmele*. Exista posibilitatea ca la constructia unei sintagme sa nu participe exact derivarea cuvântului care apare în text, în acest caz utilizându-se derivarea potrivita din familia acestuia. Referirile ulterioare la aceasta lista se vor face prin utilizarea termenului FAMLIST.
c) *Lista verbelor ce se vor transforma în substantive si vor participa la compunerea sintagmelor* se aleg în functie de domeniul stabilit. Din lista fac parte verbele si toate conjugarile acestora, precum si substantivele asociate. Procesul de transformare a verbelor în substantive mai este cunoscut si sub numele de **substantivare**. În exemplul – "Firma **a construit** cladirea în parametri proiectati." verbul **a construit** se transforma în substantivul **constructie** si apoi participa la constructia sintagmelor. Referirile ulterioare la aceasta lista se vor face prin utilizarea termenului VERBLIST.
d) *Lista cuvintelor care vor fi omise când se analizeaza textul* contine prepozitii, conjunctii, interjectii, cuvinte cu frecventa mare de aparitie ce nu sunt interesante pentru constructia sintagmelor. (Vezi explicatiile de la algoritmul de constructie a sintagmelor si de indexare a acestora.) Referirile ulterioare la aceasta lista se vor face prin utilizarea termenului STOPLIST.
e) *Operatorii* sunt identici cu cei de la regasirea dupa cuvinte indexate, singura diferenta fiind ca nu se mai aplica cuvintelor ci sintagmelor. În plus, apar doi noi operatori, **SPEC(sintagma),** pentru aflarea primei sintagme specifice, si **GEN(sintagma),** pentru aflarea primei sintagme generice.

f) *Ierarhiile arborescente care contin sintagme aflate în relatia generic-specific*. Este prezentata în continuare o astfel de ierarhie:

Activitati economice
Prestari servicii
Reparatii auto
Publicitate
Constructii
Constructii cladiri
Realizare instalatie electrica

Conform ierarhiei anterioare sintagma **Prestari servicii** este SPECIFICA pentru sintagma **Activitati economice**, iar pentru sintagma **Publicitate** este GENERICA.

g) *Algoritmul de constructie a sintagmelor si de indexarea a acestora.* Constructia sintagmelor este asistata de calculator. Exista o multitudine de algoritmi de constructie a sintagmelor, în continuare fiind prezentat unul dintre acestia. Textul care se analizeaza se desparte în fraze. Fiecare fraza, la rândul sau, se desparte în cuvinte, ignorându-se cuvintele din STOPLIST. În prima etapa se pastreaza numai substantivele ce exista în WORDLIST si numai verbele din VERBLIST, precum si ordinea de aparitie a acestora. Prin procesul de substantivare, verbele se transforma în substantive. Cu substantivele astfel obtinute se genereaza combinari de la 1 la câte cuvinte sunt, tinându-se cont de ordinea de aparitie în text. La generarea sintagmelor participa toate derivarile din familia cuvintelor pastrate anterior. Daca printre sintagmele astfel obtinute se regasesc sintagme din ierarhii, atunci acestea se leaga automat de text. Restul sintagmelor obtinute prin generare, dar care nu s-au regasit în ierarhii, sunt prezentate unui utilizator specializat care, printr-un program care îl asista, le poate insera în ierarhiile de sintagme, astfel legându-le automat la textul respectiv. Toate regulile de fiabilitate, enumerate la algoritmul de regasire dupa cuvinte indexate, trebuie sa fie respectate si în acest caz. Luând în considerare tot exemplul de mai înainte – "Firma **a construit** cladirea în parametri proiectati." – se obtin substantivele **firma, constructie** (din verbul **a construit**), **cladire, parametri, proiect**. S-a presupus ca toate cuvintele se regasesc în WORDLIST. Asa cum se observa, se utilizeaza formele de baza din familiile de cuvinte. Desi în realitate se genereaza combinari cu toate formele din familiile de cuvinte, mai jos se vor enumera numai o parte dintre cele realizate cu formele de baza.

Combinari de un cuvânt:
FIRMA
CONSTRUCTIE (se genereaza si CONSTRUCTII – sintagma candidat formata dintr-un singur substantiv care va fi prezentata operatorului, care o poate asocia sau nu cu textul din care s-a obtinut)
CLADIRE
PARAMETRI
PROIECT

Combinarile de 2 cuvinte:
FIRMA CONSTRUCTIE (se genereaza si FIRMA CONSTRUCTII, sintagma candidat care va fi prezentata operatorului, care o poate asocia sau nu cu textul din care s-a obtinut)
FIRMA CLADIRE
FIRMA PARAMETRI
FIRMA PROIECT
CONSTRUCTIE CLADIRE (se genereaza si CONSTRUCTII CLADIRI, care se regaseste în ierarhia din exemplul anterior si este automat asociata cu textul din care s-a obtinut)
PARAMETRI PROIECT
……

Combinari de 3 cuvinte:
FIRMA CONSTRUCTIE CLADIRE (se genereaza si FIRMA CONSTRUCTII CLADIRI, sintagma candidat care va fi prezentata operatorului, care o poate asocia sau nu cu textul din care s-a obtinut)
…….

Combinari de 4 cuvinte:
FIRMA CONSTRUCTIE CLADIRE PARAMETRI
…….

Combinari de 5 cuvinte:
FIRMA CONSTRUCTIE CLADIRE PARAMETRI PROIECT
…….

Limita pâna la care algoritmul genereaza combinari, în cazul în care sunt prea multe cuvinte în fraza, trebuie sa se poata seta prin aplicatie. Astfel se reduce numarul sintagmelor fara sens semantic. Tot pentru rafinarea criteriilor de generare a sintagmelor se pot utiliza cuvintele de legatura. De exemplu, sintagma FIRMA CONSTRUCTII este putin fortata si poate fi înlocuita cu sintagma FIRMA **DE** CONSTRUCTII.

h) *Regulile de actualizare* ramân aceleasi ca la regasirea dupa cuvinte indexate, doar ca se aplica sintagmelor si nu cuvintelor.

Oricare dintre conceptele enuntate anterior, atât în prima parte cât si în cea de-a doua pot suferi modificari, acestea dorindu-se doar niste repere în scrierea unui motor de cautare.

Bibliografie

[1] *Michael W. Berry, Murray Browne* – Understanding Search Engines: Mathematical Modeling and Text Retrival (Software, Enviroments Tools)

[2] *Robert M. Losce* – Text Retrival and Filtering - Analytic Models of Performance