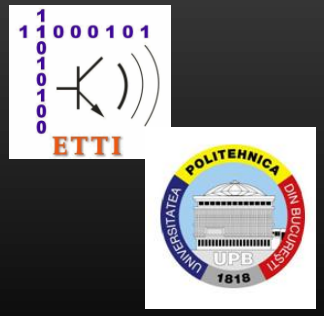


UNIVERSITATEA POLITEHNICA BUCURESTI
FACULTATEA DE ELECTRONICA, TELECOMUNICATII SI TEHNOLOGIA INFORMATIEI

WEBUL SEMANTIC

Masterand: Cocoru Vlad-Cosmin
IISC

CUPRINS

1. Scurt istoric
2. Introducere
3. Webul hipertext si limitarile sale
4. Premisele Webului Semantic
   A. Stocarea informatiilor semantice
   B. Formate de stocare (RDF si OWL)
   C. Aplicatii capabile de exploatarea informatiilor semantice (Extragerea limbii probabile, Stemming, Masuri de similaritate, Determinarea domeniului tematic al unui document, Identificarea subiectului si a personajelor, Rezumatizarea automata)
5. Concluzii

SCURT ISTORIC

1989 – Tim Berners Lee realizeaza prima comunicare HTTP server – client. Ia nastere WWW-ul (sistem global de documente hipertext interconectate logic prin hiperlegaturi). Apare familia de protocoale TCP/IP – faciliteaza comunicarea dintre sistemele de calcul din internet

1991 – 2003 – WWWul este referit retroactiv drept Web 1.0. Este caracterizat de pagini statice HTML, folosirea frameset-urilor,a guestbookurilor online si a messageboardurilor

2003 – prezent – maturizarea Web 2.0. Este caracterizat de aplicatii web ce faciliteaza raspandirea de informatii interactiv, interoperabilitate, design ce pune utilizatorul in prim plan si colaborarea dintre oameni. Vizitatorii unui site pot coopera la intretinerea informatiilor (Enciclopedii, Twitter/ Facebook, aplicatii web de tipul e-activities, retele sociale, siteuri de sharing video sau foto, bloguri)

CE URMEAZA

Se asteapta aparitia Web 3.0, care presupune dinamizarea totala a paginilor web si implementarea webului semantic, ca unealta a calculatoarelor de a observa adevaruri in text si a genera noi adevaruri (sau informatii).

Tim Berners-Lee descrie webul semantic drept o componenta a webului 3.0:

> „Oamenii ma intreaba ce este Webul 3.0. Cred ca daca adaugi grafica vectoriala – tot continutul sa fie dinamic, interactiv si atragator – peste Webul 2.0 si oferi acces la Web Semantic peste un spatiu enrom de date, obtii acces la o resursa incredibila” (T.B-L, 2006)

CE INSEAMNA SEMANTIC

Semantic = „semnificatie” sau „studiul semnificatiilor si schimbarilor acestora”.

Webul semantic implica faptul ca semnificatiile datelor sa fie descoperite nu numai de oameni, dar si de calculatoare. Prin software, acestea ar putea gasi, citi, intelege si utiliza datele din WWW precum oamenii.

Primul pas in acest sens a fost realizat prin XML care faciliteaza modul de formulare a informatiilor si de partajare a acestora. (XML – limbaj de caracterizare simpla a datelor)

Urmatorul pas este utilizarea tehnicilor de reprezentare a cunoasterii pentru exprimarea semanticii Webului si realizarea inferentelor. Abordarile utilizate in inteligenta artificiala, bazele de date si limbajele de programare trebuie reexaminate, adaptate si extinse pentru a putea cuprinde necesitatile Webului semantic

Pe scurt, trebuie adaugata logica actualului Web. Termenul utilizat este de fapt „ontologie”, adica specificatii explicite ale conceptelor (reprezentarea cunostintelor cu ajutorul unui set de concepte inter-relationate).

WEBUL HIPERTEXT SI LIMITARILE SALE

Informatiile de pe un computer tipic se impart in 2 categorii: documente (emailuri, rapoarte, brosuri – citite de oameni) si date (calendare, playlisturi, spreadsheeturi – destinate interpretarii computerizate, necesita prelucrare inaintea vizualizarii de catre om)

WWW-ul este bazat pe documente scrise in HTML, o conventie de marcare ce este folosita pentru a coda continutul unui text intretesut cu obiecte mutlimedia (imagini sau formulare interactive, de exemplu). Tagurile metadata, de exemplu:

<**meta** name = „keywords” content = „computing, computer studies, computer”>
<**meta** name=”description” content=”catalog de soliutii software pentru recunoastere de forme”>
<**meta** name=”author” content=”Vlad Cocoru”>

, reprezinta o metoda prin care calculatoarele pot categorii continutul paginilor Web.

Cu un browser web aceste documente se pot interpreta pentru a citi informatiile. Pe baza tagurilor, un interpretor poate face rationamente simple, la nivel de document, de genul „acest document este un curs despre computers

WEBUL HIPERTEXT SI LIMITARILE SALE

Limitarea este faptul ca nu se poate intra in continutul efectiv al cursului pentru a culege informatiile.

Nu existe capacitati specifice HTML pentru a afirma ca la pozitia 5 se gaseste, de exemplu, cel mai popular pachet software de recunoastere de forme. Nici nu se poate face legatura dintre acest item si pretul lui, cum nu se poate face nici distinctia intre un item din lista si preturile aferente.

PREMISELE WEBULUI SEMANTIC

Pentru a putea vorbi de un Web Semantic este nevoie de 3 elemente:

A. Stocarea informatiilor semantice
B. Aparitia de formate de stocare a datelor semantice
C. Existenta de aplicatii web capabile sa opereze si sa exploateze informatiile semantice

A. STOCAREA INFORMATIILOR SEMANTICE

Se poate face fie la nivel local al paginii, fie in alte locatii web.

- **Tagging / Labelling** – procedee de evidentiere a unor elemente din continutul unei pagini web. Taggingul se refera la etichetarea unei informatii scurte, concise, pe cand labellingul presupune etichetarea unui ansamblu de informatie (cu structuri de metadate).

  Metoda are avantajul puterii prin simplitate, pregatind informatia pentru clasificari

- **Uniform Resource Identifier (URI)** – string de identificare ce poate fi de natura unei locatii (URL) sau a unui nume (URN). Asemanari: URN – nume persoana, URL – adresa persoana. In esenta, este tot o forma de etichetare a unor elemente de continut web, dar permite articularea unui limbaj comun intre utilizatori (prin referire la obiecte clar identificate prin ID unic).

De exemplu:

- pagina web non-semantica – descriere item: **<item>**calculator**</item>**
- Pagina web cu continut semantic **<item rdf:about**=”http://database.org/resource/computer”**>**calculator**</item>**, aratand unui browser semantic unde sa caute informatii despre obiectul pe care il reprezinta.

B. FORMATE STOCARE A DATELOR SEMANTICE

B1. RDF (RESOURCE DESCRIPTION FRAMEWORK)

Modelul RDF este similar conceptului de modelare sub forma de diagrame de clase. Se bazeaza pe ideea de a construi propozitii despre resurse (Web) sub forma unor structuri sintactice de forma Subiect – Predicat – Obiect (terminologia RDF: *„triple”*)

Subiectul denota sursa, iar predicatul denota trasaturi ale resursei ce indica relatia dintre subiect si obiect. De exemplu informatia „Cerul este de culoare albastra” este pusa sub forma unei triple RDF:

Subiect - „cerul” **Predicat** - „are culoarea” **Obiect** - „albastru”

Adevarurile din baza de date pot fi modelate sub forma unor propozitii ce suporta refomrulari, fiecare afirmatie (tripla RDF) primind un URI si fiind tratata ca o resursa edspre care se pot face alte afirmatii. De exemplu „Alina spune ca *cerul este de culoare albastra”.*

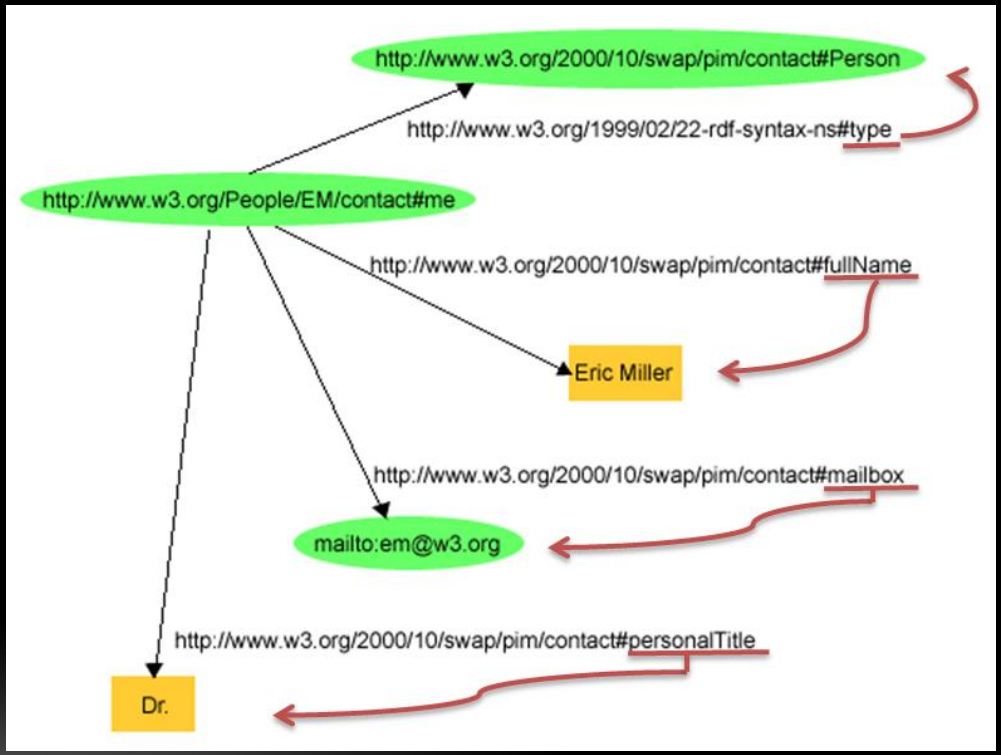
Rezultatul: fiecare element din tripla poate fi la randul lui o alta afirmatie (identificata prin URI). Astfel se creeaza legaturi complexe intre elemente.

B. FORMATE STOCARE A DATELOR SEMANTICE

B1. RDF (RESOURCE DESCRIPTION FRAMEWORK)

Exista implementari ale modelului RDF care propun ideea de a grupa propozitiile pe baza unor criterii, numite „situatii” sau „contexte”. O propozitie poate fi asociata cu un context (referit prin URI) cu scopul de a obtine o informatie de genul *„este adevarata in contextul X”.*

O alta propunere este retinerea sursei fiecarei propozitii, identificabila tot prin URI. Cand se modifica sursa, afirmatiile modelului care au fost deduse din acea sursa se pot modifica dupa caz.



Exemplu de graf rezultat din RDF

B. FORMATE STOCARE A DATELOR SEMANTICE

B2. OWL (WEB ONTOLOGY LANGUAGE)

Familie de limbaje pentru reprezentarea cunostintelor si descrierea ontologiilor pentru Web. 2004 – introducere OWL, 2009 – OWL 2.0. Interes militar, didactic, medical.

Datele sunt interpretate ca un set de elemente ce au o serie de „afirmatii despre proprietate”, care relationeaza elementele intre ele. Axiomele ce se deduc (care caracterizeaza legaturile intre elemente - numite clase - si specifica tipurile de relatii dintre elemente), ofera inteles semantic, permitand sistemelor sa formuleze informatii aditionale pe baza datelor prezentate explicit.

De exemplu, o ontologie ce descrie familiile poate include axiome ce afirma ca proprietatea „AreMama” este prezenta intre 2 indivizi numai cand exista si proprietatea „AreParinte”, si ca indivizii clasei „AreGrupaSanguina0” nu vor fi niciodata legati prin proprietatea „AreMama” de indivizi „AreGrupaSanguinaAB”.

B. FORMATE STOCARE A DATELOR SEMANTICE
B2. OWL (WEB ONTOLOGY LANGUAGE)

Altfel spus, daca se afirma ca Alex este legat prin proprietatea „AreMama” de Maria, si ca Alex este un membru al clasei „AreGrupaSanguina0”, se poate afirma ca Maria nu este un membru al clasei „AreGrupaSanguinaAB”

OWL opereaza asupra claselor de entitati URI, stabilind relatii generice intre ele. Se pot genera automat RDF-uri, adica adevaruri noi.

Emitand axiome si reguli de constructie logica, se pot genera constructii logice, de tipul teoremelor intalnite in gramaticile formale.

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA
INTRODUCERE

De exemplu, browserele semantice, capabile sa ofere cautari inteligente, ce folosesc principiile sinonimiei si ale stemmingului, clusterizari semantice ale raspunsurilor, capacitatea de a face sugestii de cautare sau sumarizari automate.

O alta categorie ar fi aplicatiile ce nu au ca scop primar cautarea de informatii on-line, ci ghidarea unui utilizator uman in desfasurarea unor activitati, si oferirea de sfaturi pe care aplicatia le descopera prin intermediul webului semantic (aplicatia intelege contextul in care lucreaza utilizatorul uman si ofera sugestii si cauta solutii pe baza informatiilor semantice).

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C1. EXTRAGEREA LIMBII PROBABILE

Un cuvant stop este un termen specific unei limbi, care are urmatoarele 2 calitati: este foarte frecvent folosit in acea limba si nu poarta o amprenta tematica bine definita. Daca un text contine un numar suficient de mare de cuvinte stop ale unei anumite limbi, este de presupus ca este scris in limba respectiva.

Densitatea de utilizare a cuvintelor stop, si in general a oricaror cuvinte intr-un text, se exprima sub forma frecventei de utilizare la suta de cuvinte:

F= (Nr. Incidente)/(Nr Total Cuvinte) *100.

Decizia asupra limbii probabil a textului este o problema care presupune existenta unei limite. Fie **T** un text si **F** frecventa de utilizare a cuvintelor stop dintr-l limba data **L**. Experimental, se determina un prag **p** pentru care

T** este scris in limba **L** daca **F>p

Blocurile decizionale pot fi monolingvistice sau polilingvistice si pot actiona la nivelul intregului text sau la nivelul paragrafelor sau chiar a propozitiilor, permitand astfel o corecta tratare a documentelor neuniforme din punct de vedere lingvistic.

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C1. EXTRAGEREA LIMBII PROBABILE

I	are	certainly	entirely	gave	how's	matter	nothing
a	aren't	chosen	equally	general	I'd	maybe	now
aboard	around	clear	especially	generally	if	me	occasionally
about	as	clearly	etc	get	I'll	mean	of
above	aside	common	even	getting	I'm	means	off
according	ask	completely	eventually	given	immediately	meant	on
across	at	constantly	ever	giving	in	merely	once
actually	away	could	every	gone	indeed	might	one
additional	back	couldn't	everybody	got	instead	mightn't	only
after	be	daily	everyone	greatly	into	mine	onto
again	became	daren't	everything	had	is	Miss	or
against	because	dawn	everywhere	hadn't	isn't	more	or
ago	become	determine	exactly	halfway	it	most	other
ahead	becoming	did	extra	happen	its	mostly	ought
alike	been	didn't	fact	has	it's	must	oughtn't
all	before	do	far	hasn't	itself	mustn't	our
almost	behind	does	farther	have	i've	my	ours
alone	being	doesn't	feature	haven't	just	myself	ourselves
along	belong	doing	few	having	kind	naturally	out
already	below	done	fewer	hay	larger	near	outer
also	beneath	don't	fifteen	he	last	nearby	outline
although	beside	double	fifth	he'd	late	nearer	outside
am	best	down	fifty	he'll	least	nearest	over
among	better	dozen	five	hello	less	nearly	own
amount	between	during	follow	her	let	necessary	partly
an	beyond	each	for	here	let's	needed	parts
and	both	earlier	forth	here's	like	needn't	perfectly
another	brief	early	forty	hers	likely	never	perhaps
any	built	easily	four	herself	lot	next	possibly
anybody	but	eight	fourth	he's	made	nine	previous
anyone	by	either	frequently	him	mainly	no	probably
anything	cannot	eleven	from	himself	make	nobody	proper
anyway	can't	else	full	his	making	none	properly
anywhere	carefully	enough	fully	how	many	nor	put
apart	certain	entire	further	however	March	not	putting

Cuvinte stop din limba engleza (fragment)

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C2. STEMMING

Stemmingul este o tehnica de normalizare a multimilor cuvintelor unei limbi in clase de echivalenta, in cadrul carora termenii sa fie inruditi morfologic si semantic. Desi seamana cu procesul lexical de extragere a radacinii cuvintelor, nu este acelasi lucru, intrucat accentul nu cade pe identificarea radicinii si sensului acestuia, ci pe semnatica.

Extragerea radacinii unui cuvant (desufixizare, deprefixizare). De exemplu:

infrigurat -> in-frig-ur(i)-ate => frig (lat. frigus / frigoris)
frigorific -> frig-orific => frig (lat. frigus / frigoris)
frigaruie -> frig(e) ar(e)-uie => frige(lat. frigere)

Stemming -deprefixizarea este absenta intrucant produce intotdeauna schimbari ale sensului: Operatia de stemming pe baza exemplelor de mai sus:

infrigurat -> frigur-ate => infrigur
frigorific -> frigor-ific => frigor
frigaruie -> frigar-uie => frigar

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C2. STEMMING

Pentru algorimtii de stemming (automat) se intocmeste o lista a sufixelor si regulilor de desufixizare aferente Sufixele se elimina in ordinea descrescatoare a lungimii lor.

Un algoritm de stemming pentru limba romana:

Pas 0: eliminarea formelor de plurar (si alte simplificari). Cautarea celui mai lung sufix dintre urmatoarele si efectuarea operatiei indicate:

ul ului	-> stergere
aua	-> inlocuire cu ***a***
ea ele elor	-> inlocuire cu ***e***
ii iua iei iile iilor ilor	-> inlocuire cu ***i***
ile	-> inlocuire cu ***i*** daca nu este precedat de ***ab***
atei	-> inlocuire cu ***at***
aţie aţia	-> inlocuire cu ***aţi***

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA
C2. STEMMING

Pas 1: reducerea sufixelor combinative. Cautarea celui mai lung dintre sufixele urmatoare si efectuarea operatiei indicate. Repetarea acestui pas pana cand nu se mai face nicio modificare:

abilitate abilitati abilităi abilităţi	-> inlocuire cu ***abil***
ibilitate	-> inlocuire cu ***ibil***
ivitate ivitati ivităi ivităţi	-> inlocuire cu ***iv***
icitate icitati icităi icităţi icator icatori iciv iciva icive icivi icivă ical icala icale icali icală	-> inlocuire cu ***ic***
ativ ativa ative ativi ativă aţiune atoare ator atori ătoare ător ători	-> inlocuire cu ***at***
itiv itiva itive itivi itivă iţiune itoare itor itori	-> inlocuire cu ***it***

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA
C2. STEMMING

Pas 2: eliminarea sufixelor standard. Cautarea celui mai lung sufix dintre urmatoarele, si efectuarea operatiei indicate:

at ata ată ati ate ut uta ută uti ute it ita ită iti ite ic ica ice ici ică abil abila abile abili abilă ibil ibila ibile ibili ibilă oasa oasă oase os osi oşi ant anta ante a nti antă ator atori itate itati ităi ităţi iv iva ive ivi ivă

-> sterge

iune iuni -> sterge daca este precedat de un ***ţ***, si inlocuieste ***ţ*** cu ***t***.

ism isme ist ista iste isti istă işti -> inlocuieste cu ***ist***

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C2. STEMMING

Pas 3: (doar daca nu s-a eliminat niciun sufix la pasii 1 si 2) Eliminarea sufixelor verbale. Cautarea celui mai lung sufix dintre urmatoarele si efectuarea actiunii indicate.

are ere ire âre indând indu ându eze ească ez ezi ează esc eşti eşte ăsc ăşti ăşte am ai au eam eai ea eaţi eau iam iai ia iaţi iau ui aşi ărăm ărăţi ară uşi urăm urăţi ură işi irăm irăţi ară âi âşi ârăm ârăţi âră asem aseşi ase aserăm aserăţi as eră isem iseşi ise iserăm iserăţi iseră âsem âseşi âse âserăm âserăţi âseră usem useşi use userăm userăţi useră -> sterge daca este precedat de o consoana sau de ***u***

ăm aţi em eţi im iţi âm âţi seşi serăm serăţi seră sei se sesem seseşi sese seserăm seserăţi seseră -> sterge

Pas 4: eliminarea vocalei finale. Cautarea celui mai lungi dintre urmatoarele sufixe si eliminarea lui:
a e i ie ă

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C3. MASURI DE SIMILARITATE

In urma analizei unui text din caruia i s-a determinat limba, s-au eliminat cuvintele stop si s-a aplicat algoritmul de stemming, se poate face un calcul al frecventei utilizarii cuvintelor. Rezulta o posibila plasare vectoriala a textului in spatiul limbii respective, daca fiecare cuvant (rezultat in urma stemmingului) este considerat a fi o axa a spatiului, iar fiecare frecventa este considerata coordonata in functie de acea axa.

Majoritatea componentelor unui astfel de vector de frecvente ar fi nule, iar dimensiunea spatiului imaginat este de regula extrem de mare, de ordinul sutelor de mii sau milioanelor. Totusi, cu puterea de calcul actuala, aceasta nu reprezinta un inconvenient.

Avand 2 texte **T1** si **T2**, si reprezentarile lor vectoriale **F1** si **F2**, se poate imagina o masura de similaritate ca fiind :

- norma euclidiana **DE(T1,T2) = || F1 – F2 ||** sau
- produsul scalar: **DS(T1,T2) = <F1-F2>**

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C4. DETERMINAREA DOMENIULUI TEMATIC AL UNUI DOCUMENT

Aceasta abordare presupune efectuarea a 2 pasi premergatori:

1) alegerea unui numar de domenii care sa acopere realitaeta discursurilor in general, la o rezolutie cat mai potriivta. De exemplu, s-ar putea propune lista:

Dom = { Administratie, Mediu, Finante, Auto, Economie, Educatie, Imobiliare-Constructii, IT&C, Media, Politica, Moda, Sanatate, Justitie, Sport, Stiinta, Turist, Vedete-Monden, Sex, Cultura, Animale, Meteo} (***cu 21 de elemente***)**.**

2) Adunarea de documente bine incadrate tematic, pentru fiecare dintre domeniile din lista, si intr-un numar suficient de mare. Pentru fiecare document se va face o radiografie a notiunilor comune, a denumirilor si expresiilor, adica se va intocmi un vector al frecventelor de utilizare.

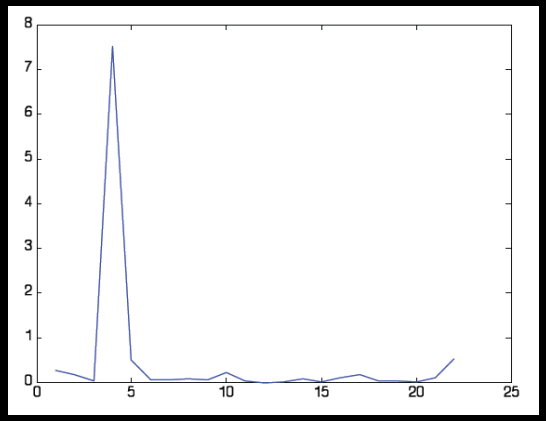
Apoi, la analiza unui text se poate face analiza termenilor si plasarea fiecarui cuvant intr-unul din domeniile disponibile, cu o anumita probabilitate. In final, se va face o medie a tuturor cuvintelor si se poate spune cu o probabilitate destul de ridicata la ce domeniu se refera documentul.

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

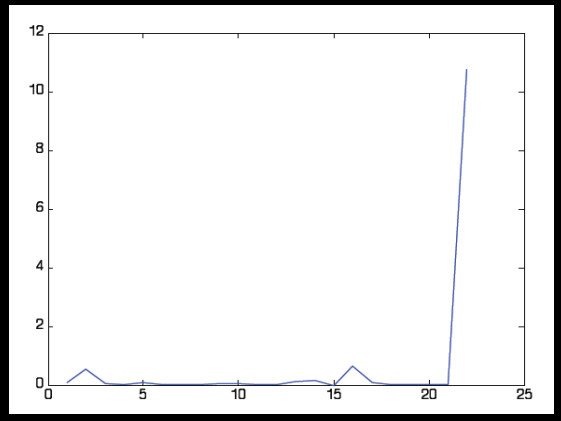
C4. DETERMINAREA DOMENIULUI TEMATIC AL UNUI DOCUMENT

TERMEN	DOMENIU 1	DOMENIU 2	DOMENIU 3	DOMENIU M
constant	0.01	0.2	0	0.11
construct	0.3	0	0	0.3

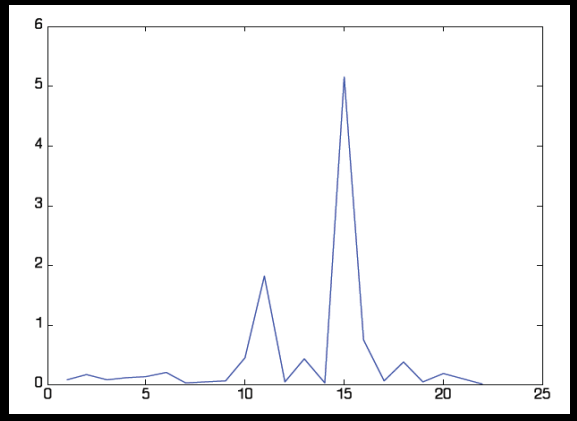
Exemplu de rezultat al calculului probabilitatilor domeniilor unui cuvant



Profil tematic al cuvantului „masina”, cu varful pe domeniul „Auto”



Profil tematic al cuvantului „grad”, cu varful pe domeniul „Meteo”



Incadrare tematica a unui text ce relateaza despre un parlamentar european ce detine o echipa de fotbal.

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C5. IDENTIFICAREA SUBIECTULUI SI A PERSONAJELOR

Personajele, sintagmele si citatele joaca intr-un text rolul reperelor. Ele permit o rapida orientare in legatura cu sensul si incadrarea tematica a unui discurs. De aceea, pentru o abordare automata a unui document, este necesara dezvoltarea unor instrumente ce recunosc si extrag aceste componente.

Personajele se pot recunoaste:

- **autarhic** – prin logica interna a unui document si a limbii. Are avantajul ca permite descoperirea unor personaje noi si flexibilitatea si autonomia in functionare, dar are dezavantajul unei rate ridicate de confuzii si erori, induse de indisciplina sintactica si ortoepica a documentelor.
- **In mod catalog** – pe baza unor dictionare externe de personaje. Avantajul prezentat este corectitudinea extragerilor, dar dezavantajul este ca nu permite identificarea de termeni noi.

Similar, se poate construi o lista de abrevieri. Se cunoaste ca in functie de context, o abreviere inseamna lucruri diferite.

Exemplu: ITI = *International Theater Institution (Unesco)* sau *Information Technology Institute* sau *Intestinal Tract Infection*. Pe baza determinarii contextului se poate inlocui un acronim cu varianta sa completa.

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C6. REZUMATIZARE AUTOMATA

Rezumatul unui text coerent este un alt text coerent, mai scurt, care reuseste sa comprime informatiile primului. Nu exista un standard, dar exista o orientare „intuitiva” a rezumatizarii corecte. Elementele ce trebuiesc regasite intr-un rezumat de calitate sunt:

A. Concizia absoluta

B. Concizia relativa la un standard de lungime impus

C. Cursivitatea logica a textului

D. Pastrarea doar a personajelor centrale

E. Pastrarea doar a povestii centrale

F. Repovestirea in limbaj minimalist

Rezumatele sunt destinate lecturii umane. Niciunul din criteriile de mai sus nu poate fi apreciat decat in mod subiectiv, de aceea calitatea rezumatizarii nu poate fi masurata propriu-zis. Cum insa criteriul **A** este extrem de relativ, intrucat nu depinde doar de subiectivitate, cat si de scopul urmarit si chiar de aspecte legate de copyright, iar **F** presupune cunoasterea si recunoasterea sintaxei, pentru a putea elimina figurile de stil si propozitiile secundare din cadrul unei fraze, gasirea formularilor alternative, se propune implementarea unui algoritm care sa respecte criteriile **B**, **C**, **D** si **E**.

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C6. REZUMATIZARE AUTOMATA

Pas 1: Descompunerea textului in propozitii/fraze. In functie de algoritm, se pot obtine rezolutii diferite, de la identificarea propozitiilor simple (dictate de logica gramaticala, sintactica si de punctuatie) pana la identificarea doar a frazelor (dictata de regulile de punctuatie)

T -> { P1, P2, .. , Pn}

Pas 2: Se extrag toate cuvintele din **T** si se echivaleaza cu clasa lor stemmica, se traseaza chestiunea termenilor proprii (identificare, reducere la formele eliptice, sinonimizari) si a sintagmelor, cat si numararea aparitiilor la nivel de label stemm. Se acorda o valoare de importanta fiecarei notiuni, in functie de natura si frecventa de utilizare in intreg textul. Reguli de notare posibile:

- Un nume propriu este mai important ca o sintagma
- O sintagma este mai importanta ca un cuvant comun
- Un termen este mai important daca frecventa de utilizare este mai ridicata in text
- Un termen este mai important daca frecventa de utilizare este mai ridicata in limba
- Un termen este mai important daca are un profil tematic mai asimetric in limba
- Un termen de la inceputul unui text e mai important ca unul de la mijlocul textului
- Un termen de la sfarsitul textului este mai important ca unul de la mijlocul textului

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA
C6. REZUMATIZARE AUTOMATA

In urma acestui pas rezulta un tabel al importantelor cuvintelor, de forma:

Termen	carte	Ion Creanga	Amintiri din copilarie	...
Importanta	0.01	0.12	0.7	...

Pas 3: Se realizeaza o punctare a fiecarei propozitii in parte, in care se folosesc informatiie de la pasul anterior. Propozitiile primesc un scor prin insumarea ponderata a valorilor de importanta a cuvintelor care le compun.

Propozitia	P1	P2	...	Pn
Scorul	s1	s2	...	sn

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA

C6. REZUMATIZARE AUTOMATA

Pas 4: Se realizaeaza o matrice a distantelor de asemanare semantica intre propozitii (pe baza masurilor de similaritate). Cu ajutorul acestor distante se modifica scorul fiecarei propozitii. Daca o propozitie este legata semantic de alte propozitii, va primi un plus de scor de la fiecare, proportional cu scorul propozitiei surori si invers proportional cu distanta semantica fata de aceasta.

Astfel, se ofera posibilitatea ca unele propozitii de legatura, ce nu aveau un scor important, sa reintre in joc, facand textul mai citibil

Pas 5: Se reindexeaza propozitiile, in ordinea descrescatoare a scorului calculat anterior. Informatia se completeaza cu informatii de gabarit (nr cuvinte, nr caractere) asupra fiecarei propozitii si se retine sub forma unui tabel de genul:

Propozitia	P1	P2	...	Pn
Scorul	S1 >	S2 >	... >	Sn
Indexul original	i1	i2	...	in
Nr. cuvinte	w1	w2	...	wn
Nr. caractere	c1	c2	...	cn

C. APLICATII WEB CE FOLOSESC INFO SEMANTICA
C6. REZUMATIZARE AUTOMATA

Pas 6: Se extrage rezumatul. Gabaritul rezumatului poate fi controlat la diferite rezolutii:

- Procent din numarul de propozitii originale
- Procent din numarul de cuvinte originale
- Procent din numarul de caractere initiale
- Numar aproximativ de cuvinte
- Numar aproximativ de caractere

Dupa setarea gabaritului, se colecteaza in limitele lui propozitiile cu scorurile cele mai mari. Apoi aceste propozitii se reordoneaza dupa indexul original si se asambleaza intr-un text.

CONCLUZII: CE INSEAMNA WEB SEMANTIC

- **Date accesibile calculatoarelor** – webul semantic presupune ideea de a avea datele pe web definite si legate in modalitati care sa permita calculatoarelor sa le utilizeze nu numai in scopul afisarii lor, dar mai ales pentru automatizarea, integrarea si utilizarea lor in diferite aplicatii

- **Agenti inteligenti** – scopul webului semantic este sa faca webul actual mai usor de citit de calculatoare, pentru a permite agentilor inteligenti artificiali sa restabileasca si aa manipuleze informatii

- **Baze de date distribuite**: webul semantic sa faca pentru date ceea ce HTML-ul a facut pentru informatiile textuale. Sa poata sa reprezinte toate bazele de date si regulile logice ale acestora ca un tot. Este o incercare de a face pentru datele prelucrabile de catre calculatoare ceea ce WWWul a facut pentru documentele citibile de catre oameni.

- **Infrastructura informatizata**: webul semantic va fi o infrastructura, nu o aplicatie. Problema este ca webului ii lipseste o structura de baza usor de automatizat.

CONCLUZII: CE INSEAMNA WEB SEMANTIC

- **Serviciu pentru umanitate**: aplicatiile vor putea sa elibereze utilizatorul uman de sarcina consumatoare de timp de a localizare resursele relevante pe web, sa le extraga, sa integreze si sa indexeze informatiile continute in ele. Se doreste colectarea automata de informatii din diferite surse, integrarea lor si procesarea lor pentru extragerea adevarurilor, pe care apoi aplicatiile le vor prezenta utilizatorilor drept produs finit.

- **Adnotare superioara**(cu metadate): webul semantic va asigura adnotare evoluata pentru documente, in formate accesibile procesarii automate si efectuarii legaturilor dintre ele

- **Cautare imbunatatita**. Dezideratul este accesarea informatiilor de pe web avand in vedere continutul, si mai putin cuvintele.

&