Ceaușescu Ciprian - Mihai
Gupa 406 – Inginerie Software
Facultatea de Matematică și Informatică

Adnotarea informației

-Referat-

Ceaușescu Ciprian - Mihai

Grupa 406 Inginerie Software

Ceaușescu Ciprian - Mihai
Gupa 406 – Inginerie Software
Facultatea de Matematică și Informatică

Cuprins

Ceaușescu Ciprian - Mihai
Gupa 406 – Inginerie Software
Facultatea de Matematică și Informatică

Adnotarea informației

1. Introducere

Una din cele mai importante invenții din ultimele decenii o reprezintă Serviciile Web. Acestea au rolul de a împărtăși informațiile între aplicații folosind internetul. Recent au apărut diverse puncte slabe, cum ar fi parcurgerea informațiilor fără a lua în considerare semnificația lor. Acestea determină nevoia de a apărea un nou Web care să aibă mai multă relevanță pentru utilizator.

Semantic Web este o extensie a Web-ului actual deoarece reprezintă informațiile cu mai mult sens atât pentru oameni cât și pentru calculatoare. Permite descrierea conținutului și serviciilor într-o formă care poate fi înțeleasă de mașini și permite automatizarea proceselor precum adnotarea, descoperirea, publicitatea, publicarea și compunerea serviciilor. Web-ul Semantic permite descrierea formală a resurselor existente pe Internet (pagini Web, documente text şi multimedia, baze de date, servicii etc). Dintre avantajele acestuia se impune ca principală: identificarea rapidă și precisă a resurselor relevante pentru utilizator precum şi exploatarea automată a resurselor de către agenţii inteligenţi. Ideea de Web Semantic a apărut în urmă cu aproximativ 15 ani şi a fost introdusa de către Tim Berners-Lee, inventatorul Web-ului. Această formă a Web-ului permite calculatoarelor să utilizeze informațiile de pe web în același mod în care o fac oamenii. Cu alte cuvinte, semantica informațiilor este bine definită în web-ul semantic pentru a face posibilă extragerea informațiilor. Totuși, web-ul conține informații distibuite și eterogene. De exemplu, să presupunem că doriți să aflați mai multe informații despre un profesor care predă informatică la o universitate din apropiere. L-ați văzut la o conferință acum două luni și vă mai amintiți doar numele lui de familie. Cu structura actuală a web-ului este greu să găsiți astfel de informații, deoarece acestea sunt distribuite pe pagini web diferite. Totuși este ușor să aflați ce aveți nevoie cu ajutorul web-ului semantic.

A fost dezvoltat având la bază ontologia, considerată fiind ca o “coloană vertebrală” a Web-ului semantic. Una din funcțiile Web-ului este aceea de a construi o sursă de referințe pentru diferite subiecte, în timp ce Web-ul semantic este conceput pentru a construi un Web cu sens. Ontologia reprezintă fundamentul vocabularului și al comunicării în Web-ul semantic.

Ceaușescu Ciprian - Mihai
Gupa 406 – Inginerie Software
Facultatea de Matematică și Informatică

2. Definirea adnotațiilor

Adnotarea poate fi definită ca procedura de adăugare a datelor despre date în interiorul unui document. Aceste date noi sunt denumite metadate care se traduc literalmente "date despre date". Acest conținut adăugat, poate să se refere la orice nivel al unui anumit document, cum ar fi un cuvânt, o expresie, un paragraf sau întregul document. Adnotarea poate fi considerată ca fiind o altă modalitate de etichetare a documentelor. Ambele au scopul de a îmbogăți datele, oferind astfel informații despre datele din interiorul unor documente. Cu toate acestea, adnotarea este diferită de etichetare, prin faptul că aceasta nu numai că oferă o modalitate de creștere a performanței căutării, dar oferă în plus și o modalitate de găsirea a rezultatelor care nu sunt legate de interogarea de căutare.

Adnotarea unei anumite parți a unui document poate ajuta la creșterea vitezei cu care informațiile sunt găsite sau pentru la procesul de căutare a unui anumit element. Un alt avantaj al adnotării este reprezentat de faptul că poate conecta cuvinte dintr-un document cu alte concepte stocate într-o baze de date, într-o ontologie sau pe alte suporturi de stocare a informațiilor. În același timp, adnotarea poate oferi o modalitate de clasificare a documentelor. Este destul de evident faptul că adnotarea are multe avantaje.

Procesul de adnotare poate fi de două tipuri. Primul este **adnotarea manuală** și al doilea este **adnotarea automată**.

Adnotarea manuală se bazează pe sprijinul unor persoane diferite în vederea furnizării acestor adnotări. Avantajul adnotării manuale este reprezentat de faptul că aceasta este una precis realizată când este făcută de un personal calificat, dar pe de altă parte necesită mult timp și poate fi foarte costisitor. În primul rând, persoana care se ocupă de crearea adnotării trebuie să citească întregul document pentru a obține o idee generală despre conținut, iar în această fază nu se face nici o adnotare. După aceasta, documentul este recitit și fiecare apariție a unei entități este marcată. Următorul pas care trebuie realizat este acela în care adnotarea este revizuită pentru a elimina orice greșeală. În ultimă instanță, orice decizie care s-a dovedit dificil de făcut sau orice incertitudine privind o adnotare trebuie introdusă și salvată într-un document, pentru ca o a doua persoană să verifice dacă s-au luat deciziile corecte privind adnotările realizate.

Acest proces nu este foarte eficient din punct de vedere al resurselor, deoarece necesită foarte mult timp pentru persoana care crează adnotarea. De asemenea, pentru că este nevoie de intervenția factorului uman, etapele realizate sunt predispuse la erori, deoarece, contează cât de familiarizat este adnotatorul cu domeniul acesta, dar și volumul de pregătire personală pe care l-a avut acesta.

Pe de altă parte, adnotarea automată, se bazează pe ajutorul unor instrumente diferite de creare a metadatelor. Cel mai mare avantaj al utilizării acestui tip de adnotare este faptul că este eficient în ceea ce privește volumul de muncă, pentru că se pot gestiona mai multe documente decât o persoană care încearcă să realizeza aceeași adnotare. Totuși, deoarece procesul este automatizat și lipsește orice interacțiune umană se poate prezenta, de asemenea, ca fiind predispusă la erori. Un compromis între cele două abordări este acela de a le

combina pe ambele într-un sistem automat de adnotare. În acest tip de abordare, procesul de creare a metadatelor începe cu câțiva pași de adnotare manuală, în care se extrage o listă de cuvinte cheie din document. Pe baza acestei liste, programul de adnotare automată poate găsi cuvintele care sunt similare cu acestea, prin căutarea într-un set mare de documente. Următorul pas este filtrarea acestor rezultate într-un mod manual sau automat. Avantajul acestui tip de abordare este acela că, adnotarea automată combină eficiența adnotării manuale cu viteza celei automate, reducând costurile de timp, dar și minimizând foarte mult rate de eroare.

Adnotările resurselor Web pot fi create tradițional utilizând instrumente de adnotare a documentelor sau abordări mai recente, cum ar fi Sematics Wiki, Semantics blogs și etichetare colaborativă. Momentan nu există un model unificat pentru toate aceste tipuri diferite de adnotări, fiind dificil atât să se compare și să se evalueze instrumentele de adnotare, cât și să se integreze diferitele tipuri de date de adnotare. Vom analiza în lucrarea de față tipurile de adnotări din diferite domenii și vom prezenta un model unificat pentru adnotările semantice. Vom evalua, de asemenea, instrumentele de adnotare existente în aceste domenii diferite și vom arăta cum să cartografiem datele acestor instrumente pentru modelul nostru formal, permițând astfel reprezentarea datelor într-un mod unitar.

Termenul "adnotare" implică, într-un mod foarte general, atașarea datelor unor alte date. în cursul acestei lucrări, vom elabora acest lucru printr-o referire destul de simplă la o serie de domenii diferite. Următoarele secțiuni vor oferi o scurtă introducere în fiecare dintre aceste domenii și se va specifica rolul adnotărilor pentru fiecare domeniu.

2.1. Adnotarea documentelor

Domeniul tradițional al **adnotării documentelor** acoperă adnotarea arbitrară a documentelor text sau a diferitelor părți ale acestora. Adnotările pot fi manuale (efectuate de către una sau mai multe persoane), semi-automat (pe baza sugestiilor automate), sau complet automat. Instrumentele de adnotare manuală permit utilizatorilor să adauge adnotări paginilor web sau altor resurse, și să le ofere aceste informații altor persoane. O exemplu de adnotare ar lega textul "Paris" la o ontologie, identificând-o ca oraș și capitală a Franței. Automat, instrumentele pot efectua adnotări similare (cum ar fi recunoașterea entității numite) fără intervenție manuală.

2.2. Semantic Wikis

Semantic Wikis sunt medii de creare a hipertextului colaborativ, care permite oamenilor colectarea, descrierea și scrierea informațiilor în mod colectiv. Semantic Wikis permite utilizatorilor să facă descrieri oficiale ale resurselor prin adnotarea paginilor pe care le reprezintă aceste resurse. În timp ce un Wiki obișnuit permite utilizatorilor să descrie resursele într-un limbaj natural, un Semantic Wiki le permite utilizatorilor să descrie resurse suplimentare într-o limbă oficială. Prin adăugarea de metadate la conținutul unui Wiki obișnuit, utilizatorii

beneficiază de avantaje suplimentare, cum ar fi îmbunătățirea regăsirii informației, schimbul de informații și reutilizarea cunoștințelor.

2.3. Semantic Blogs

Blogurile (sau jurnalele web) sunt reprezentate de reviste sau jurnale online. Blogurile sunt alcătuite din postări individuale, create și prezentate în ordine cronologică inversă.

O adnotare în aceste bloguri este, cel mai adesea, o declarație despre o postare. Spre exemplu, multe soluții actuale pentru a crea un blog, permit clasificarea postărilor cu simple categorii sau subiecte cum ar fi "sport", "cinema" sau "Sigmund Freud" (o formă de etichetare, discutată în continuare) - putem spune astfel că postările pe blog sunt adnotate cu aceste categorii. Într-un semantic blog aceste adnotări sunt extinse și permit asocierea din punct de vederea ontologic.

2.4. Tagging

Etichetele exprimă unele relații nespecificate dintre resursă și orice termen care se referă la aceasta. De exemplu, în Flickr o fotografie a unei pisici poate fi marcată cu "pisică", indicând că fotografia descrie o pisică. În mod similar, în del.icio.us, site-ul pentru ISWC 2006 poate fi marcat cu "semantics conference iswc2006", indicând faptul că site-ul web este despre conferința ISWC și despre semantica web.

3. Crearea adnotațiilor

Putem organiza adnotările într-un spațiu tridimensional, caracterizat după cum urmează: **efortul de adnotare**, **completitudinea rezultatului** (adică cât de bine captează situația din lumea reală) și **angajamentul** (ontologic sau social) față de rezultat. De exemplu, etichetele necesită puțin efort (sunt ușor de atribuit) și rezultă într-un angajament ridicat (prin procesul de etichetare colaborativă a comunității, aceasta este de acord cu rezultatele obținute), dar au o completitudine scăzută (nu se pot face declarații complexe despre lumea reală, fiind atribuite doar etichete superficiale).

3.1. Modelul conceptual

Termenul **adnotare** poate reprezenta atât procesul de adnotare, cât și rezultatul acestui proces. În cazul în care ne referim la o **adnotare**, ne vom referi la rezultatul acesteia. O adnotare oferă informații sub forma de date unor alte date. Cu alte cuvinte, stabilește într-un anumit context, o relație (scrisă) între datele adnotate și datele de adnotat. Distingem trei tipuri

de adnotări: **informale**, **formale** și **ontologice**. Adnotările informale nu pot fi interpretate de o anumită mașină, deoarece nu utilizează pentru definirea acesteia un limbaj formal. Adnotările formale sunt ușor de înțeles de mașină, dar nu folosesc termeni ontologici. În cazul adnotărilor ontologice, terminologia indică faptul că acestea corespund unui concept comun, numit ontologie. Un termen este definit ca fiind ontologic într-o manieră socială, diferită de cea tehnică sau formală. Este uneori înțeles greșit faptul că folosirea unui limbaj ontologic formal duce la apariția unor termeni ontologici. O ontologie, denotă însă o înţelegere comună (socială). Rezumând, putem astfel distinge trei tipuri de adnotări:

1. **adnotări informale**

2. **adnotări formale**, care au constituenți definiți într-un mod formal și care sunt astfel ușor de citit

3. **adnotări ontologice**, care au constituenți definiți formal și folosesc numai termeni ontologici, care sunt acceptați și înțeleși din punct de vedere social.

3.2. Modelul formal

Analizând mai în detaliu, putem modela adnotările sub forma unui cvadrupul, după cum urmează:

Definiția 1 (adnotare). O adnotare A este reprezetată de un tuplu (a_s, a_p, a_o, a_c), unde a_s este subiectul adnotării (datele adnotate), a_o este obiectul adnotării (datele de adnotare), a_p este predicatul (relația de adnotare) care definește tipul relației dintre a_s și a_o, și a_c este contextul în care este realizată adnotarea.

Subiectul adnotării poate fi **formal** sau **informal**. De exemplu, când punem o notă în marginea unui paragraf, convenția informală este dată de faptul că nota se aplică paragrafului, dar acest indicator nu este definit într-un mod formal. Dacă totuși vom folosi o abordare formală, cum ar fi un URI, pentru a indica paragraful, se observă faptul că subiectul este specificat într-un mod formal.

Predicatul adnotării poate fi formal sau informal. De exemplu, când punem o notă în margine, relația nu este definită în mod oficial, dar putem deriva acest lucru din contextul în care nota este un comentariu, o cerere de schimbare, o aprobarea sau o dezaprobare, etc. Dacă folosim un indicator formal pentru un termen ontologic, care indică relația (de exemplu, dc: comentariu), atunci predicatul este definit într-un mod formal.

Obiectul adnotării poate fi formal sau informal. Dacă un obiect este formal putem distinge diferite nivele de formalitate: **textuale**, **structurale** sau **ontologice**. Spre exemplu, șirul "Este minunat!" este un obiect textual. Un calcul al bugetului în marginea unei propuneri de proiect este un obiect structural. O adnotare care nu este structurală în mod explicit, ci folosește și termeni ontologici este un obiect ontologic.

Contextul de adnotare poate fi formal sau informal. Acesta ar putea indica când s-a făcut adnotarea și de către cine, sau în ce scop adnotarea realizată este considerată valabilă. De exemplu, într-un scop temporal (este valabilă numai în 2006) sau într-un domeniu de activitate care se ocupă cu spațiul (este valabil numai în Europa de Vest). Adesea, contextul este oferit într-un mod informal și implicit. Dacă folosim un pointer formal, cum ar fi un URI atunci contextul este definit formal.

Definiția 2 (adnotare oficială). O adnotare formală A_f este o adnotare A, unde subiectul este un URI, predicatul a_p este un URI, obiectul a_o este un URI sau un context formal și contextul a_c este un URI.

Definiția 3 (adnotare ontologică). O adnotare ontologică A_s este o adnotarea formală A_f, unde predicatul a_p și contextul a_c sunt elemente ontologice, iar obiectul a_o corespunde unei definiții ontologice a lui a_p.

Figura 1: adnotare informală

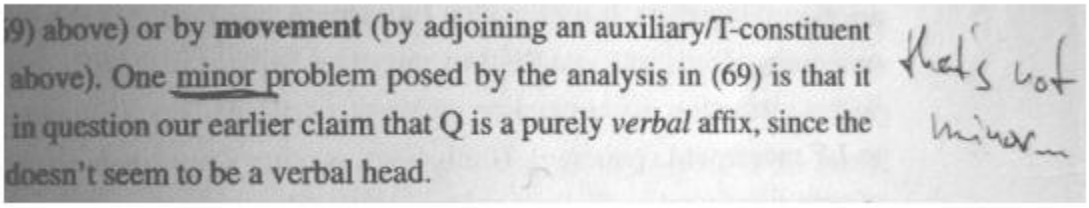

9) above) or by **movement** (by adjoining an auxiliary/T-constituent above). One minor problem posed by the analysis in (69) is that it in question our earlier claim that Q is a purely *verbal* affix, since the doesn't seem to be a verbal head.

Figura 2: adnotare formală

```
<http://papers.org/minimalism#minor>
 <disagree> "that's not minor!".
```

Figura 3: adnotare ontologică

```
<http://papers.org/minimalism#minor>
 ibis:con
  [ rdf:type ibis:Argument;
    rdf:label "that's not minor!" ].
```

Ceaușescu Ciprian - Mihai
Gupa 406 – Inginerie Software
Facultatea de Matematică și Informatică

4. Exemple de adnotații

Diferitele exemple în care adnotațiile se folosesc în practică au scopuri diferite, dar și moduri diferite. Următoarea listă de exemple nu este exhaustivă:

1. **Bibliografiile adnotate** - oferă descrieri ale modului în care fiecare sursă este utilă unui autor în construirea unei lucrări sau a unei teze.

2. **Adnotarea lingvistică** - se poate referi la orice date descriptive sau analitice aplicate datelor neprelucrate care există într-o anumită limbă. Aceste adnotări sunt adesea textuale (de exemplu, transcrieri), dar pot include și orice alt tip de date (de exemplu, analiză sintactică). O colecție de texte cu adnotări lingvistice este cunoscută ca un „**corpus**”.

3. Programatorii adesea adaugă **adnotări** de tipul comentariilor codului sursă pe care îl dezvoltă, explicând într-o manieră mai ușor de citit de om funcția codului și, probabil, idei de dezvoltare.

4. Studiile asupra **adnotațiilor ADN-ului** sunt în desfășurare încă din anii 1980 în domeniul biologiei moleculare. Aceste adnotații reprezintă de obicei câmpuri predefinite în bazele de date secvențiale.

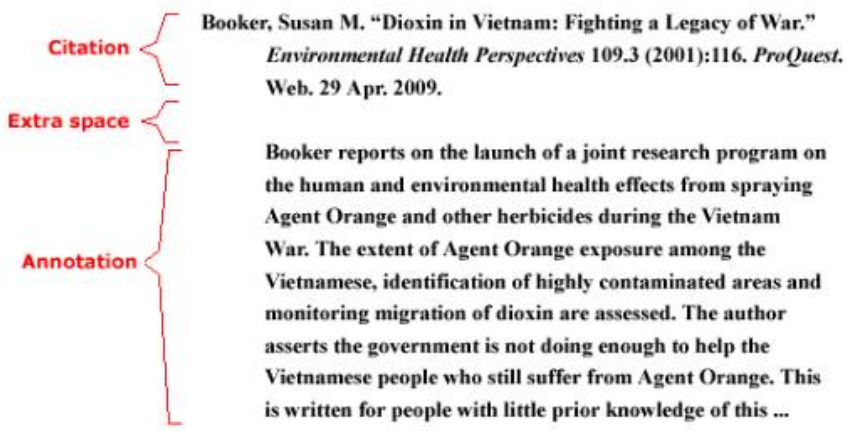
5. **Adnotarea automată a imaginii** este procesul prin care un sistem informatic atribuie metadate sub formă de cuvinte cheie unei imagini digitale. Imaginile pot fi, de asemenea, adnotate cu informații tehnice care accelerează recuperarea informației bazată pe conținut, de exemplu histograme ale conținutului de culoare și informațiilor de segmentare.

6. Artiștii de toate felurile realizează adnotări obiectelor fizice (de exemplu scoruri și scripturi) în timpul procesului de repetiție pe care îl desfășoară. Înregistrările digitale pot fi, de asemenea, adnotate pentru a crește înțelegerea mesajului transmis, spre exemplu, putem considera **adnotarea video,** care poate fi sub formă de timbre, text sau imagini suprapuse, alături de diverse comentarii.

7. Adnotările pot fi asociate cu o resursă web și pot permite utilizatorilor să personalizeze o pagină web (pentru editarea, adăugarea sau eliminarea de informații de pe pagină) fără a modifica resursele de bază. Datorită conectivității lor, **adnotările web** sunt într-o strânsă legătură. Multe aplicații web utilizează adnotări extinse și imaginative pentru a revizui, a evalua, a îmbunătăți, a adapta, a discuta, a califica sau a critica conținutul oferit atât de proprietarii site-ului, cât și de alți utilizatori.

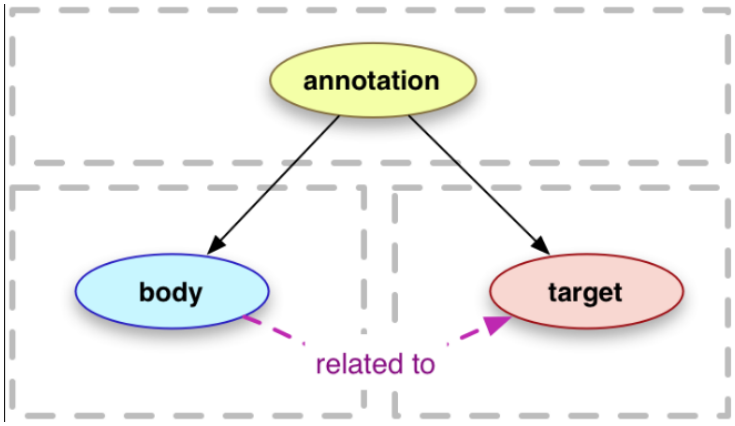
8. **Adnotarea colaborativă** utilizează etichete gratuite pentru cuvintele cheie cu scopul de a clasifica conținutul, astfel încât poate crea o taxonomie generată de utilizatori, cunoscută sub numele de **folksonomie**. Acest tip de **„social bookmarking”** este popular datorită faptului că participarea este rapidă și intuitivă și poate duce la noi modalități de vizualizare a legăturilor dintre informații, cum ar fi „tag clouds”.

Exemplul 1:



Exemplul 3:

Exemplul 5:



5. Beneficiile adnotărilor

Adnotările descriptive, ca de exemplu, metadatele, sunt esențiale pentru recuperarea informațiilor. De exemplu, **adnotările simple** de cuvinte cheie pot să rezume un text lung sau să descrie materiale non-textuale care să permită utilizarea instrumentelor de căutare a textului pentru a descoperi imagini, audio sau video. Multe sisteme de regăsirea a informației bazate pe conținut utilizează, de asemenea, adnotări non-textuale, care distribuie diferite aspecte ale fișierelor în date care pot fi procesate mai rapid, de exemplu muzica poate fi regăsită comparând datele numerice extrase anterior cu nivelurile de volum și imaginile poate fi regăsite prin similitudinea bazată pe histograme ale informațiilor de culoare din acestea.[1]

Adnotările sporesc, de asemenea, datele de bază prin furnizarea de explicații și interpretări. În acest fel, fiecare contribuție oferă viitorilor utilizatori date îmbogățite prin mai multe opinii și mai multe domenii de expertiză. Adnotările pot chiar să identifice erorile din datele de bază.

În științe, cercetarea devine din ce în ce mai dependentă de reutilizarea seturilor de date existente. Ca parte a unei baze de cunoștințe în continua dezvoltare, adnotarea este un instrument crucial de învățare și interpretare. În unele domenii, cum ar fi biologia moleculară, valoarea principală constă mai degrabă în adnotări decât în datele de bază, iar aceste informații reprezintă o investiție imensă de abilități și efort.[2]

Adnotările pot ajuta la integrarea rezultatelor obținute de diferite grupuri de cercetare prin furnizarea de informații de agregare pentru mai multe baze de date și prin identificarea provenienței fiecărei părți a datelor. Unul dintre scopurile primare ale adnotării este de a disemina informații (de exemplu, în critică textuală sau cartografie). Cu toate acestea, adnotările oferă, de asemenea, informații suplimentare despre modul în care este utilizată o resursă, despre persoanele care o utilizează și despre contextul contribuției acestora. În acest fel, adnotările susțin noi cercetări cu un accent mai larg decât cele ale datelor de bază în sine, cum ar fi cercetarea comportamentelor sociale de colaborare prin procesul de adnotare.[3]

Adnotarea acceptă colaborarea, fie că este vorba de o echipă mică cu expertiză similară, fie că este deschisă contribuțiilor oricărei persoane care are o conexiune la Web. De asemenea, poate promova o implicare sporită a subiectului. Instrumentele automate de adnotare reprezintă o provocare de cercetare în sine și pot economisi o cantitate mare de timp și efort din adăugarea manuală de informații suplimentare utile. Acest domeniu de cercetare este important atât pentru instituțiile comerciale, cât și pentru instituțiile de cercetare.[4]

Ceaușescu Ciprian - Mihai
Gupa 406 – Inginerie Software
Facultatea de Matematică și Informatică

7. Concluzii

În concluzie, adnotarea informațiilor poate fi definită ca procedura de adăugare a datelor despre date în interiorul unui document. Aceste date noi sunt denumite metadate care se traduc literalmente "date despre date". Acest conținut adăugat, poate să se refere la orice nivel al unui anumit document, cum ar fi un cuvânt, o expresie, un paragraf sau întregul document.

Adnotarea poate fi considerată ca fiind o altă modalitate de etichetare a documentelor. Ambele au scopul de a îmbogăți datele, oferind astfel informații despre datele din interiorul unor documente. Cu toate acestea, adnotarea este diferită de etichetare, prin faptul că aceasta nu numai că oferă o modalitate de creștere a performanței căutării, dar oferă în plus și o modalitate de găsirea a rezultatelor care nu sunt legate de interogarea de căutare.

Ceaușescu Ciprian - Mihai
Gupa 406 – Inginerie Software
Facultatea de Matematică și Informatică

8. Bibliografie

1. "Information Retrieval" - DigiCULT Technology Watch Report 3, December 2004

2. Buneman P. et al. "Annotation in Scientific Data: a Scoping Report"

3. Tags Networks Narrative - http://www.ioct.dmu.ac.uk/tnn/

4. For example: Matellanes, A. et al. "Creating an application for automatic annotation of images and video" , Automatic Annotation and Information Retrieval : New Perspectives (Special Track at the 20th International Florida Artificial Intelligence Research Society Conference 2007); Wyman, S. et al. (2004). "Automatic annotation of organellar genomes with DOGMA"

5. Annotation:

   http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/annotation#c4

6. Reeve Lawrence, Han Hyoil: Survey of Semantic annotation paltforms

7. Passin Thomas: Explorer Guide to the Semantic Web