

Rezumatul fazei de executie

Asa cum s-a precizat in capitolul 2, obiectivul principal al acestei faze de executie consta in elaborarea modelelor experimentale sisteme si integrarea, verificarea si validarea acestora.

In cadrul acestei etape s-au elaborat (i) modelul experimental al sistemului de extragere a informatiilor si adnotare semantica, (ii) modelul experimental al sistemelor de rationare si regasire avansata a informatiilor si (iii) modelul experimental al sistemului integrat.

Pentru elaborarea modelului experimental al sistemului de extragere a informatiilor si adnotare semantica am definit (i) o metoda de extragere a informatiilor relevante din documentele text, (ii) o metoda de adnotare semantica bazata pe reguli JAPE pentru identificarea relatiilor de tip instance-of, subclass-of si a relatiilor generale, (iii) o metoda de imbogatire a ontologiei bazata pe framework-ul Text2Onto pe care l-am adaptat in contextul nostru, (iv) o arhitectura a modelului experimental al sistemului de extragere a informatiilor si adnotare semantica. In procesului de extragere a informatiilor si adnotare semantica s-a pornit de la o ontologie core la care au fost adăugate noi concepte, în urma studiului a 100 de documente arhivistice, ajungându-se astfel la o ierarhie de concepte istorice mai completă. Ontologia core de domeniu a fost îmbogățită semi-automat, prin aplicarea metodei de achiziție de cunoștințe din documente arhivistice în format text, ce adaptează Text2Onto pentru a se potrivi cerințelor specifice limbii române. Astfel s-au utilizat și adaptat tehnologii de procesare lingvistică și de adnotare lexicală și semantică de documente. În plus, pe baza adnotărilor obținute, au fost extrase primitive de modelare cu ajutorul algoritmilor existenți în Text2Onto, a celor adaptați sau a celor noi creați. Ulterior, primitivele au fost translate într-un limbaj standard de reprezentare a ontologiilor, și anume OWL. Pentru persistarea rezultatelor, ontologia care se obține în urma procesării documentelor este stocată în baza de date ArhiNet, pentru a putea fi accesată de toate modulele sistemului. Tehnologiile și resursele necesare în metoda de achiziție de date au fost adaptate și pentru a permite prelucrarea de texte scrise cu diacritice. Această realizare a permis obținerea de rezultate mai corecte, datorită faptului că, în limba română, diacriticele influențează caracteristicile lexicale și sintactice ale cuvintelor din conținutul documentelor. Metodele de extragere a informatiilor si adnotare semantica au fost validate de ontologia rezultată în urma procesării a 115 de documente istorice furnizate de Arhivele Naționale din Cluj, referitoare la trecutul istoric al

Transilvaniei. Din procesarea acestor documente s-a obținut un număr de aproximativ 53 de concepte, 608 de instanțe și 47 de proprietăți. Ontologia astfel obținută nu este completă, având în vedere faptul că în cadrul documentelor nu sunt specificate toate detaliile despre termenii întâlniți în conținutul textelor. În consecință, este necesară o revizuire a rezultatelor obținute de către un expert de cunoștințe, pentru a avea o ontologie cât mai completă. Acest lucru poate fi realizat datorită modulului de gestionare a ontologiei, prezent în cadrul sistemului ArhiNet. Modificările efectuate de expert sunt actualizate și în modelul probabilistic de ontologie.

Modelul experimental al sistemului de rationare este responsabil de urmatoarele operatii: verificarea consistentei ontologiei, clasificarea conceptelor ontologice, realizare, inferenta bazata pe reguli si procesarea interogarilor in SPARQL. Modelul experimental al sistemului de rationare propune stocarea partiala a ontologiei de domeniu intr-o baza de date relationala din ratiuni de scalabilitate. Astfel, cunostintele de tip TBOX vor fi stocate intr-un fisier OWL care poate fi procesat de un rationator OWL-DL, iar cunostintele de tip ABOX vor fi stocate intr-o baza de date relationala. Aceasta stocare separata a cunostintelor necesita dezvoltarea unor noi metode de rationare deoarece cele existente nu se pot utiliza pe ontologii separate astfel. Astfel, regulile SWRL devin proceduri stocate SQL si interogarile SPARQL devin interogari SQL. Pentru a optimiza procesul de inferenta SWRL asupra bazei de date, continutul acesteia a fost impartit in doua tabele una continand tripletele nou inserate, iar cea de-a doua continand tripletele vechi, iar datele inferate sunt limitate la implicatii asupra tripletelor nou adaugate. Optimizarea este rafinata prin gruparea atomilor astfel incat un numar exponential de interogari este evitat. In plus, daca exista seturi de atomi care apar in mai mult de o regula SWRL, datele asociate acestor atomi sunt partajate si reutilizate de cat mai multe reguli. Un alt aspect important este legat de procesarea interogarilor SPARQL care contin atat atomi ABox cat și atomi TBOX. Pentru tratarea acestui caz am definit o metoda care imparte interogarea intr-o interogare ABOX și una TBOX si care imbina rezultatele executarii separate a celor doua interogari. In ceea ce priveste cunostintele stocate in ontologia OWL, am definit o metoda pentru rezolvarea anomaliiilor ontologice precum inconsistentia si nesatisfiabilitatea. Modelul experimental al sistemului de rationare include si o functionalitate de gestionare a regulilor SWRL.

Modelul experimental al sistemului de regasire avansata a informatiilor pune la dispozitia utilizatorilor o interfata de interogare in limbaj natural a sistemului ArhiNet. Pentru realizarea acestui model am dezvoltat urmatoarele: (i) gramatica EBNF de interogari pentru limba romana si engleza, (ii) metoda de sugestii care mai intai tokenizeaza si parseaza interogarile in

limbaj natural dupa care identifica sugestiile potrivite pe care le ordoneaza in functie de relevanta lor, (iii) metoda de traducere a interogarilor in limbaj natural in format SPARQL, interogarile in SPARQL fiind apoi transmise sistemului de rationare pentru furnizarea rezultatelor si (iv) metoda de procesare si afisare a rezultatelor returnate de sistemul de rationare.

Modelele experimentale ale celor trei sisteme descrise mai sus au fost testate si apoi integrate intr-un sistem distribuit accesibil pe internet.

Rezultatele obtinute in cadrul acestei etape au fost diseminate in cadrul a 9 articole stiintifice publicate / acceptate spre publicare in publicatii indexate in baze de date internationale (ISI, IEEE, ACM, reviste B+ conform CNCSIS). De asemenea, in vederea diseminarii rezultatelor a fost organizat workshopul "Workshop on Digital Libraries, e-Content Management and e-Learning" in cadrul conferintei "2010 IEEE International Conference on Intelligent Computer Communication and Processing - ICCP 2010" pe data de 26 august 2010 la Cluj-Napoca.