Universitatea POLITEHNICA din București
Facultatea de Electronică, Telecomunicații și Tehnologia Informației
Universitatea Babeș-Bolyai
Facultatea de Matematică și Informatică

Romagna: o abordare modernă asupra uneltelor de analiză conceptuală a datelor

Lucrare de licență

Prezentată ca cerință parțială pentru obținerea
titlului de *Inginer*
în domeniul *Calculatoare și tehnologia informației*
programul de studii *Ingineria informației*

Conducător științific
Christian Săcarea

Absolvent
Mihai Chereji

Anul 2014

Declaraţie de onestitate academică

Prin prezenta declar că lucrarea cu titlul *Romagna: o abordare modernă asupra uneltelor de analiză conceptuală a datelor*, prezentată în cadrul Facultăţii de Electronică, Telecomunicaţii şi Tehnologia Informaţiei a Universităţii "Politehnica" din Bucureşti ca cerinţă parţială pentru obţinerea titlului de *Inginer* în domeniul Inginerie Electronică şi Telecomunicaţii/ Calculatoare şi Tehnologia Informaţiei, programul de studii *Ingineria informației* este scrisă de mine şi nu a mai fost prezentată niciodată la o facultate sau instituţie de învăţământ superior din ţară sau străinătate. Declar că toate sursele utilizate, inclusiv cele de pe Internet, sunt indicate în lucrare, ca referinţe bibliografice. Fragmentele de text din alte surse, reproduse exact, chiar şi în traducere proprie din altă limbă, sunt scrise între ghilimele şi fac referinţă la sursă. Reformularea în cuvinte proprii a textelor scrise de către alţi autori face referinţă la sursă. Înţeleg că plagiatul constituie infracţiune şi se sancţionează conform legilor în vigoare. Declar că toate rezultatele simulărilor, experimentelor şi măsurătorilor pe care le prezint ca fiind făcute de mine, precum şi metodele prin care au fost obţinute, sunt reale şi provin din respectivele simulări, experimente şi măsurători. Înţeleg că falsificarea datelor şi rezultatelor constituie fraudă şi se sancţionează conform regulamentelor în vigoare.

Bucureşti, Iulie 2014.
Absolvent: Mihai Chereji

.....

Cuprins

Lista figurilor

Lista tabelelor

Lista acronimelor

FCA = Formal Concept Analysis (Analiza Conceptuală Formală)
DOM = Document Object Model
SVG = Scalable Vector Graphics (Grafice Scalabile Vectoriale)
HTML = HyperText Markup Language (limbaj de marcaj hiper-textual)
XML = eXtendable Markup Language (Limbaj extensibil de marcaj)
SQL = Structured Query Language (Limbaj de Interogare Structurată)
LLVM = Low Level Virtual Machine (Mașină virtuală la nivel scăzut)
MVC = Model View Controller

Introducere

Big-data a devenit un domeniu foarte căutat în zilele noastre, devenit un cuvânt repetat de toată lumea, de la programatori la oameni de marketing. Acest lucru se-ntâmplă deoarece lumea se *îneacă* în date, adunate din diferite surse. Site-urile înregistrează fiecare mișcare a utilizatorilor, pedometre care încarcă pe Internet numărul de pași făcut de utilizator în fiecare oră, aparatele de analiză medicală întorc tot mai multe date.

Interpretarea lor devine din ce în ce mai grea, și tendința este de a se folosi metode de *machine learning*, ceea ce presupune învățarea automată a calculatoarelor prin prelucrarea repetată a unei cantități imense de date. .

Analiza conceptuală formală oferă o alternativă, în care adunarea datelor este automată dar prelucrarea și interpretarea datelor este realizată de om, bazându-se pe câteva principii matematice bine definite și (relativ) ușor de înțeles.

Pe scurt, analiza conceptuală formală dezvoltă o metodă care este folosită în principal pentru a analiza date. Altfel spus, prin această metodă se pot deriva relații implicite între obiecte care sunt descrise prin legături intre seturi de atribute. Conceptele se numesc așa deoarece datele sunt structurate in unitati care sa reprezinta concepte ale gandirii umane, sau abstractii formale ale acestora, pe care un om va putea sa le interpreteze cu usurinta.

Analiza conceptuală formală mai poate fi definită și ca o metodă de clasificare conceptuală (clustering conceptual).

Acest câmp a început în anii 80 la Universitatea din Darmstadt, ca o evoluție a teoriei clasice a laticelor. Între timp, s-a dezvoltat într-un câmp care adună tot mai mulți cercetători și își dovedește utilitatea practică dincolo de câmpul din care s-a tras și dincolo de matematica abstractă.

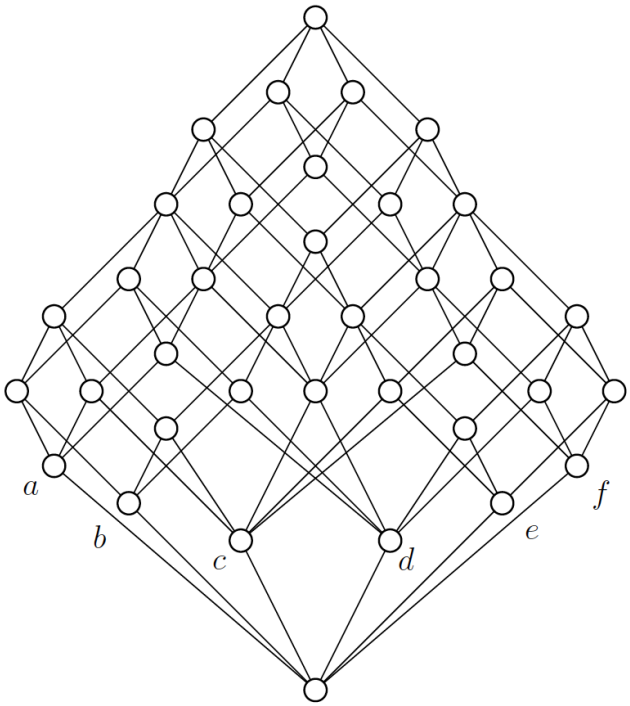
Ce e mai spectaculos e că totul se bazează pe câteva idei simple, elegante, generale, frumoase am putea spune, încât duc aproape de *filozofia platonică* (un concept este alăturarea unei mulțimi de obiecte descrise de unele proprietăți și mulțimea acelor proprietăți).

Unele din cele mai utilizate rezultate ale acestei metode de analiză sunt laticele de concept, diagrame care reprezintă ierarhiile conceptelor descrise mai sus.

Se observă că dincolo de utilitatea practică (descrisă în capitolele ce urmează), laticea este *frumoasă*. Fără a încerca să-i descifreze înțelesul, oricine poate observa că este unul din acele produse ale științei care, ne-intenționat, poate fascina și studenții artelor.

Cu toate acestea, domeniul FCA duce lipsă de unelte frumoase (vom explica, în capitolele ce urmează ce înțelegem prin *software frumos*), chiar dacă există tot mai multe programe care doresc să ajute cercetătorii în acest domeniu. Unele frustrează încă de la încercarea procesului de instalare, altele mistifică cercetătorul prin mesaje de eroare criptice, iar altele, poate, doar se mișcă prea încet, sau sunt prea greu de folosit.

În cele ce urmează dorim să descriem pe scurt domeniul analizei conceptuale formale (în capitolul 1), starea soft-ului existent acum (în capitolul 2), tehnologiile folosite la dezvoltarea



unei noi aplicații în capitolul 4 și o viziune mai în detaliu asupra noii aplicații (capitolul **??**), care dorește să rezolve unele din problemele întâlnite de oamenii interesați în acest domeniu, și poate, să atragă prin software bine gândit (și, sperăm, *bine realizat*, căci un concept fără nici un obiect nu are cuprins - *veți înțelege după ce veți parcurge primul capitol*) adepți noi.

Capitolul 1
Analiza conceptuală formală

1.1 Introducere

Analiza conceptuală formală este o metodă de sistematizare a datelor în **concepte**, definite la modul larg ca tupluri constituite din mulțimi de obiecte care împărtășesc anumite atribute și mulțimea acelor atribute. Este o reinterpretare a teoriei clasică a laticelor, dezvoltată în principal în anii '30, axată către partea practică. Conceptul a fost introdus în lucrarea seminală a lui Robert Wille din 1982 [Wil82], iar termenul a fost introdus în 1984 de același autor. În ultimele decenii, domeniul a atras multe contribuții și și-a dovedit utilitatea în domenii cum ar fi analiza și vizualizarea datelor, managementul informației.

1.2 Concepte matematice de bază

Pentru a avea o mai bună înțelegere a conceptelor pe care se bazează aplicația, vom explica pe scurt definițiile de bază ale domeniului.

1.2.1 Mulțimi ordonate, latice, latice complete

Vom începe cu câteva concepte elementare de algebră a mulțimilor, deoarece acest domeniu se bazează în întregime pe ele. Următoarele definiții sunt preluate din [GW97], înafara cazului unde este menționat altfel.

Definiție 1. *O mulțime M este ordonată dacă se poate aplica asupra sa o relație binară R care are următoarele proprietăți:*

Reflexivitate $xRx, \forall x \in M$

Antisimetrie $xRy, x \neq y \Rightarrow yRx$ *e fals,* $\forall x, y \in M$

Tranzitivitate $xRy, yRz \Rightarrow xRz, \forall x, y, z \in M$

Relația R se numește o ***relație de ordine****, sau* ***relație de ordine parțială****.*

Cel mai simplu exemplu, intuitiv exemplu este mulțimea numerelor reale $\mathbb{R}$, alături de relația $\leq$. Notăm o mulțime ordonată cu relația $\leq$ cu $(M, \leq)$. Un alt exemplu, relevant domeniului, este mulțimea tuturor submulțimilor sau *mulțimea părților* a unei mulțimi (notată cu $\mathcal{P}(M)$ pentru mulțimea M), și relația de incluziune ($\subseteq$).

Definiție 2. *Un element x al mulțimii M este* ***acoperit*** *de y dacă $x < y$ și nu există nici un z astfel încât $x < z < y$. În mod invers, x* ***acoperit*** *al lui y.*

Putem nota relația de acoperire astfel: $x \prec y, y \succ x$.

Definiție 3. *Două elemente ale unei mulțimi ordonate sunt* ***comparabile*** *(în raport cu $\leq$) dacă $x \leq y$ sau $y \leq x$ (adică relația $\leq$ se aplică asupra lor). Altfel sunt* ***incomparabile****. Un* ***lanț*** *este o submulțime în care oricare două elemente sunt comparabile. Un* ***antilanț*** *este o submulțime în care oricare două elemente sunt incomparabile.*

Definiție 4. ***Diagramele Hasse*** *sunt o reprezentare grafică a unei mulțimi ordonate. Elementele mulțimii sunt desenate ca niște cercuri. Dacă $x, y \in M$ (mulțimea descrisă de diagramă) și $x \prec y$, se desenează o linie între cele două elemente, iar elementul acoperit este desenat sub cercul reprezentând elementul care îl acoperă.*

Astfel, putem citi o diagramă Hasse în felul următor: $x < y$ dacă și numai dacă putem urmări o linie continuă de la cercul care reprezintă x la cercul care reprezintă y.

Exemplu 1. *Vom desena o diagramă Hasse reprezentând mulțimea de numere naturale $\{1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60\}$, ordonate după divizibilitate.*

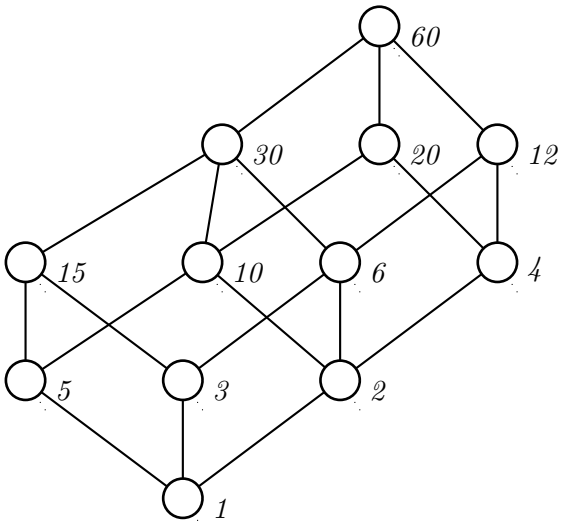


Figura 1.1: Diagramă Hasse a unei mulțimi de numere naturale, ordonate după divizibiltate.

Definiție 5. *Fie $(M, \leq)$ o mulțime ordonată, și N o submulțime a sa. Înțelegem prin* ***minorantul*** *mulțimii N un element i astfel încât $\forall a \in N, i \leq a$. În mod invers,* ***majorantul*** *mulțimii s este definit prin $\forall a \in N, s \geq a$. Putem nota mulțimea tuturor minoranților a grupului N cu I. Elementul cel mai mare din această mulțime este numit* ***infimumul*** *mulțimii N. Invers, cel mai mic element din mulțimea majoranților este numit* ***supremumul*** *mulțimii N.*

Infimumul se poate nota cu $\wedge N$ sau $\inf N$, iar supremumul cu $\vee N$ sau $\sup N$.

Definiție 6. *O mulțime ordonată M este numită o* ***latice*** *dacă $\forall x, y \in M, \exists x \vee y, \exists x \wedge y$. În alte cuvinte, o mulțime ordonată este o latice dacă pentru orice 2 elemente ale mulțimii există supremum și infimum. O latice este* ***completă*** *dacă pentru orice submulțime (finită) a ei există supremum și infimum.*

Orice latice completă are *cel mai mare element*, numit **1**, și cel mai mic element, numit **0**.

Definiție 7. *Citând din [CR04], fie o mulțime H și $\mathcal{K}$ o mulțime de submulțimi de-ale lui G. I este un sistem de închideri peste G dacă și numai dacă:*

$$\bigcap_{i\in I} A_i \in \mathcal{K} pentru\ fiecare\ submulțime\ non\text{-}vidă A_{i\in I} \subseteq \mathcal{K} și \mathcal{K} \in H$$

Astfel, $\mathcal{K}$ trebuie închisă la intersecție și să aibă un cel mai mare element. Dacă $\mathcal{K}$ e un sistem de închideri, atunci tuplul $(K, \subseteq)$ e o latice completă în care

$$\bigwedge\{A_i | i \in I\} = \cap_{i\in I} A_i,$$

$$\bigvee\{A_i | i \in I\} = \bigcap\{B \in \mathcal{K} | \cup_{i\in I} A_i \in B\}$$

Definiție 8. *Conform [CR04], un **operator de închideri** peste H e un morfism $\varphi : \mathcal{P}(H) \leftarrow \mathcal{P}(H)$ astfel încât, pentru toate $A, B \subseteq H$:*

1. $A \subseteq \varphi(A)$
2. $A \subseteq B \Rightarrow \varphi(A) \subseteq \varphi(B)$
3. $\varphi(\varphi(A)) = \varphi(A)$

[CR04] Mulțimea tuturor închiderilor al unui operator de închideri e un sistem de închideri și formează o latice completă, dacă relația de ordonare este relația de incluziune, în care:

$$\bigwedge\{A_i | i \in I\} = \cap_{i\in I} A_i,$$

$$\bigvee\{A_i | i \in I\} = \varphi\left(\bigcup_{i\ inI} A_i\right)$$

Definiție 9. *Conform [CR04] O **conexiune Galois** este compusă dintr-un tuplu de două mulțimi ordonate, M, N și două morfisme γ, ψ astfel ca $\gamma : M \to N, \psi : N \to M$, dacă și numai dacă:*

1. $m_1 \leq m_2 \Rightarrow \gamma m_1 \geq \gamma m_2$
2. $n_1 \leq n_2 \Rightarrow \psi n_1 \geq \psi m_2$
3. $m \leq \psi\gamma m, n \leq \gamma\psi n$

sau, echivalent, $m \leq \psi n \Leftrightarrow n \leq \gamma m$

În cazul unei conexiuni Galois între $\mathcal{P}(M)$ și $\mathcal{P}(N)$, morfismul $\psi\gamma$ e un operator de închidere peste M și morfismul $\gamma\psi$ e un operator de închidere peste N.

1.2.2 Context, concept, ierarhie de concepte

Definiție 10. *Fie un triplet $K = (G, M, I)$ format din 2 mulțimi, G și M, și o relație binară I între acestea. Acesta se numește **context**. Mulțimea G este compusă din obiecte, iar M din atribute.*

Literele provin din limba germană, în care aceste concepte au fost descrise inițial, fiind inițialele cuvintelor Gegenstände și Merkmale, respectiv. Relația I e numită **relația de incidență**, iar gIm poate fi citit ca "obiectul g are atributul m".

Exemplu 2. *Preluăm următorul exemplu din [CR04], un context (foarte redus)al animalelor vertebrate. Atributele sunt reprezentate pe coloane, iar obiectele sunt rânduri. Un $\times$ într-o celulă înseamnă că există relația gIm, pentru atributul și obiectul respectiv.*

		respiră în apă (a)	zboară (b)	are cioc (c)	are mâini (d)	are schelet (e)	are aripi (f)	trăiește în apă (g)	naște pui vii (h)	produce lumină (i)
1	Liliac		×			×	×		×	
2	Vultur		×	×		×	×			
3	Maimuță				×	×			×	
4	Pește papagal	×		×		×		×		
5	Pinguin			×		×	×	×		
6	Rechin	×				×		×		
7	Pește lanternă	×				×		×		×

Tabela 1.1: Un context al animalelor vertebrate. Sursa: CDA [CR04]

Pentru $A \subseteq G$, definim $A' = \{m \in M | gIm, \forall g \in A\}$.

În mod asemănător, pentru $B \subseteq M$, $B' = \{g \in G | gIm, \forall m \in B\}$.

În cuvinte, A' este mulțimea tuturor atributelor (din contextul la care ne raportăm) care descriu toate obiectele din A.

Definiție 11. *Fie $A \subseteq G$, $B \subseteq M$, unde $A' = B$ și $B' = A$. Tuplul (A, B) este un **concept** al contextului (G, M, I).*

În engleză, mulțimea A (a tuturor obiecte descrise de atributele conceptului) este numită **extent** (în română *cuprins*), iar B (atributele care descriu toate obiectele conceptului) **intent**(în română *conținut*).

Fie (G, M, I) un context, și A, A_1, A_2 submulțimi ale lui G, iar B, B_1, B_2 submulțimi de-ale lui M. Conform [GW97], atunci:

1. $A_1 \subseteq A_2 \Rightarrow A_1^{'} \supseteq A_2^{'}$
2. $A \subseteq A'''$
3. $A' = A'''$
4. $A \subseteq B' \iff B \subseteq A' \iff A \times B \subseteq I$

Proprietăți echivalente se observă imediat și pentru B, B_1, B_2.

Având în vedere că $' : \mathcal{P}(G) \to \mathcal{P}(M)$ și $B' : M \to G$, cei doi operatori pot fi combinați pentru a crea A'' și G'', care au ca domeniu mulțimea submulțimilor G și M respectiv.

Se observă pornind de la proprietățile enumerate mai sus că cele două funcții de derivare descriu o conexiune Galois între mulțimile submulțimilor pentru obiecte ($\mathcal{P}(G)$) și ($\mathcal{P}(M)$)

În exemplul de mai sus, $\{2,4\}'' = \{2,4,5\}$, $\{d,h\}'' = \{d,e,h\}$.
Câteva proprietăți de remarcat ale conceptelor, așa cum sunt definite:

- Nu orice submulțime de obiecte definește cuprinsul unui concept. Din cele descrise mai sus, rezultă că e necesar ca $A = A''$ pentru A să fie cuprinsul unui concept.

  Ca exemplu, în tabelul 1.1, $\{6\}$ nu definește un concept, deoarece $\{6\}'' = \{4,6,7\}$, adică toate atributele care descriu rechinul în contextul nostru descriu de-asemenea și peștele lanternă, și peștele papagal.

- Intersecția oricâtor cuprinsuri (sau conținuturi) de concepte are ca rezultat întotdeauna un alt cuprins (respectiv conținut).

- În urma reuniunii lor, pe de altă parte, rareori rezultă un alt cuprins.

- Mulțimea conceptelor unui context este o mulțime ordonată, dacă definim o relație de ordine în felul următor:

  Definiție 12. *Fie (A_1, B_1) și (A_2, B_2) concepte ale contextului $K = (G, M, I)$. Spunem că (A_2, B_2) este un **subconcept** al lui (A_1, B_1) (notat $(A_2, B_2) \leq (A_1, B_1)$ dacă $A_2 \subseteq A_1$. Astfel, (A_1, B_1) este **supraconceptul** lui (A_2, B_2).*

Teoremă 1. ***Teorema de bază a laticelor de concepte*** *- Fie un context (G, M, I), și o mulțime ordonată $\mathcal{C}$ $(G, M, I; \leq)$ se numește laticea de concepte a contextului, care are supremumul și infimumul descrise de:*

$$\bigwedge_{t \in T} (A_t, B_t) = \left(\bigcap_{t \in T} A_t, \left(\bigcup_{t \in T} B_t \right)'' \right)$$

$$\bigvee_{t \in T} (A_t, B_t) = \left(\left(\bigcup_{t \in T} A_t \right)'', \bigcap_{t \in T} B_t \right)$$

Exemplu 3. *Vom prezenta diagrama laticei de concepte derivate din tabelul 1.1. Se observă că supremumul are cuprins non-nul, dar infimumul da (pentru că nu există nici un animal în tabel cu toate prorpietățile). [CR04] Cum se citește diagrama? Gândindu-ne la definiția diagramelor Hasse (4), înseamnă că ne putem da seama dacă există o relație de ordine între două concepte (subconcept vs. supraconcept) dacă putem urmări o linie continuă între ele, cu supraconceptul afișat deasupra. Asta înseamnă că **toate** animalele descrise în context au atributul "au schelet" (deși asta era evident din descierea tabelului, fiind vorba de animale vertebrate). Alte deducții vizibile imediat e că nu toate animalele care au aripi pot zbura, și nu toate animalele care pot zbura sunt păsări.*

Se poate observa că diagrama de mai sus nu are etichete descriptive pentru fiecare concept reprezentat. Reprezentarea acelor informații pentru fiecare concept îngreunează diagrama și oferă multă informație redundantă, având în vedere că multe din concepte sunt subconceptele altor concepte. Totuși, cum hotărâm care etichete sunt afișate?

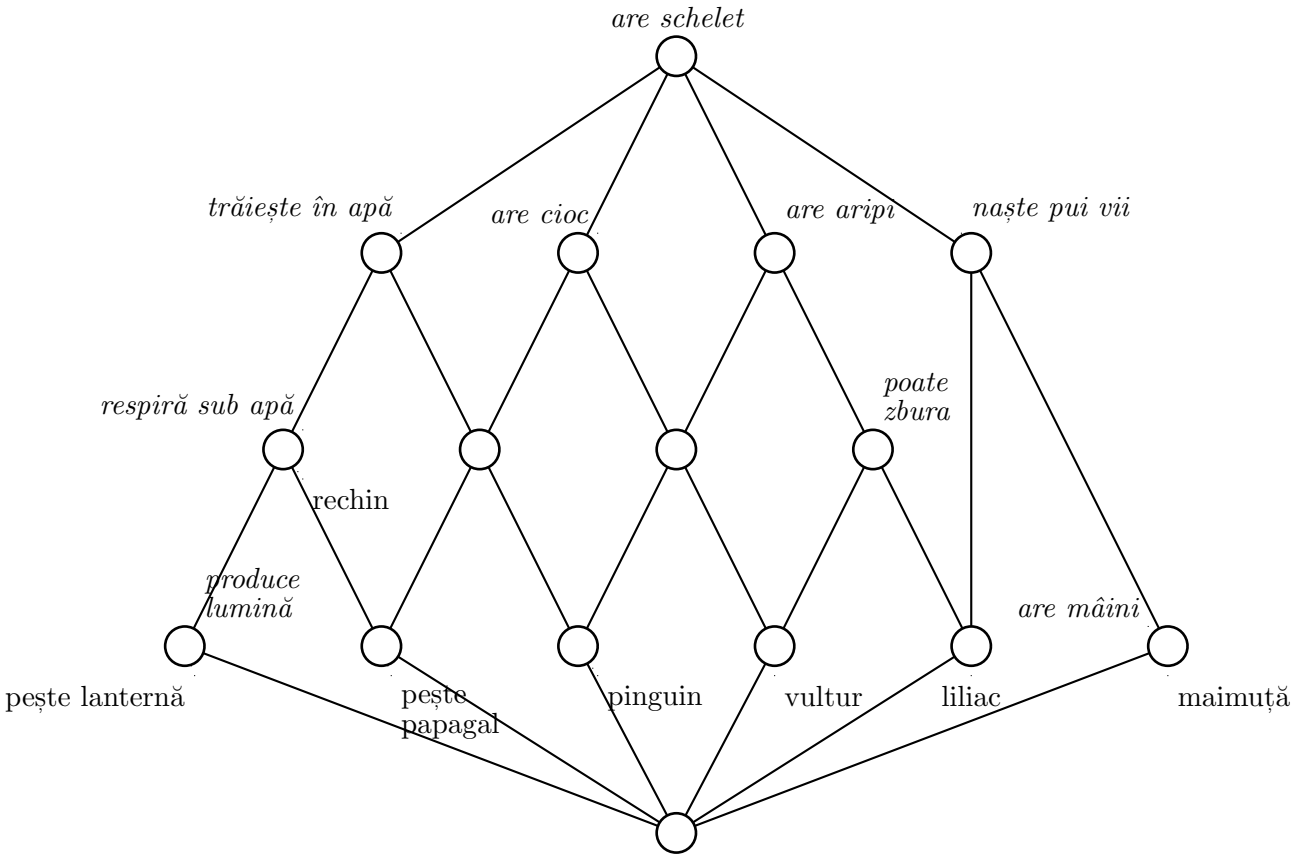


Figura 1.2: Laticea de concepte corespunzătoare contextului descris în tabelul 1.1.

Definiție 13. ***Conceptul obiectului*** $g \in G$ *este conceptul* (g'', g')*, unde* g' *este conținutul obiectului* $(\{m \in M \ gIm\})$ g*. Astfel, conținutul obiectului (notat cu* $\gamma(g)$*) este cel mai mic concept care-l are în cuprins pe* g*. Ca o paralelă,* (m'', m') *este* ***conceptul atributului*** *pentru un atribut* $m \in M$*,* m' *fiind cuprinsul atributului* $(\{g \in G | gIm\})$*. Conceptul atributului, notat cu* $\mu(m)$ *este cel mai mare concept care-l are pe* m *în conținut.*

Conform [GW97], laticile de concepte a mai multor contexte pot să fie isomorfe. Există manipulări ale contextelor pe care le putem face fără a alterastructura laticei de concepte, cum ar fi comasarea obiectelor cu același conținut și a atributelor cu același cuprins.

Definiție 14. *Fie un context* (G, M, I)*. Contextul este* ***clarificat*** *dacă* $\forall g, h \in G, g' = h' \Rightarrow g = h$ *și* $m' = n' \Rightarrow m = n, \forall m, n \in M$*.*

Laticea de concepte nu este influențată nici de prezența atributelor care pot fi scrise ca o combinație de alte atribute. Altfel spus, dacă $m \in M$ e un atribut și $X \subseteq M$ e o mulțime de atribute, iar $m \notin X$, dar $m' = X'$, atunci conceptul atributului μm e infimumul conceptelor atributelor $\mu x, x \in X$.

Altfel spus, putem omite atât atributele cât și obiectele reductibile, fără a influența laticea de concepte.

Revenind la etichetare, în loc să repetăm pentru concept care cuprinde animalul *rechin* această etichetă, putem să afișăm numele său doar pe conceptul obiectului. În loc de a îngreuna citirea diagramei, știind aceste concepte de bază diagrama devine mai lizibilă fără a fi încărcată.

- Reducerea și clarificarea contextelor
- Rezolvarea contextelor cu valori multiple

1.2.3 Contexte cu valori multiple

Contextele prezentate până acum, sunt foarte limitate prin prisma faptului că relația de incidență dintre obiect și atribut permite doar *existența* sau *lipsa* atributului pentru un anumit obiect. Dar lumea și obiectele pe care dorim să le descriem rareori pot fi descrise doar prin prezența sau lipsa unei proprietăți. Dacă descriem flori, unul din atributele de care-am vrea să ținem cont când construim conceptul este culoarea. Acest atribut nu poate fi descris simplu prin dihotomia *are/ nu are*.

Astfel, vedem nevoia de a introduce un nou concept, acela al **contextelor cu valori multiple**.

Definiție 15. *[CR04] "Fie un cvadruplu (G, M, V, I), format din 3 mulțimi G, M, V, G, mulțimea obiectelor, M, mulțimea atributelor cu valori multiple și V, mulțimea valorilor atributelor și dintr-o relație ternară între G, M și V (adică $I \subseteq G \times M \times V$) astfel încât $(g, m, v) \in I$ și $(g, m, w) \in I \Rightarrow v = w$"*

Putem citi $(g, m, v) \in I$ ca și "atributul m are valoarea v pentru obiectul g". Dar contextele cu valori multiple nu se pretează conceptualizării, cum am definit-o în secțiunea anterioară.

Pentru a putea crea concepte, contextele cu valor multiple se vor transforma în contexte simple, prin diferite metode, numite **scale**. Cea mai simplă este înlocuirea oricărui atribut cu valori multiple în tupluri formate din $M \times V$ (de exemplu pentru atributul *culoare* cu valori potențiale albastru, verde, roșu, creem atributele *culoare-albastru, culoare-verde, culoare-roșu*). Aceasta se numește o **scală nominală**.

Scalele nominale se potrivesc în cazul în care domeniul de valori al atributului pe care-l înlocuim este compus din valori care se exclud reciproc (cum e exemplul anterior, al culorilor petalelor unei flori - dacă nu luăm în considerare culorile ca fiind compuse din alte culori primare, valorile se exclud). Dar în cazul în care o valoare este subsumată de alta? Ca un exemplu, să luăm câteva valori posibile pentru atributul "luminescent": *luminos, foarte luminos, orbitor*

Se observă că primele valori posibile sunt subsumate în cele din urmă. Astfel, un bec este *luminos*, dar soarele este și *luminos*, dar și *orbitor*. Acesta este un exemplu de **scală ordinală**.

Scalele ordinale permit comparația într-un singur sens. În practică, de multe ori, vrem să aflăm domeniul în care se află o valoare, nu doar o margine inferioară sau superioară. Astfel, exemplul de mai sus poate fi reformulat cu valori cum ar fi $\leq 100lm$, $\geq 100lm$, $\leq 500lm$, $\geq 500lm$, $\leq 1000lm$, $\geq 1000lm$. Astfel, un bec care luminează cu 750 lumeni va avea atributele $\geq 100lm$, $\geq 500lm$, $\leq 1000lm$. Aceasta este o **scală intra-ordinală**.

Exemplu 4. *Pentru a demonstra scalarea, vom transforma un context cu valori multiple care se referă la becuri aflate în comerț într-un context cu valori simple, folosindu-ne de cele trei tipuri de scale descrise:*

Exemplul 4 prezintă toate cele 3 tipuri de scale pe care le-am prezentat mai sus.

1.3 Algoritmi relevanți

În secțiunea ce urmează vom prezenta câțiva algoritmi folositori pentru a prelucra contextele și laticile de concepte aferente acestora pentru a putea fi afișate.

	culoare	consum curent	luminescență
incandescent standard	caldă	100W	1690lm
bazat pe LED-uri	neutră	27W	1600lm
bec fluorescent	rece	26W	1750lm
bec cu halogen	neutră	43W	870lm

Tabela 1.2: Context cu valori multiple descriind becuri

Becuri	culoare caldă	culoare neutră	culoare rece	consum curent ≥ 30	consum curent ≥ 70	consum curent ≥ 100	luminescență $\leq 1000lm$	luminescență $\geq 1000lm$	luminescență $\leq 1300lm$	luminescență $\geq 1300lm$	luminescență $\leq 1700lm$	luminescență $\geq 1700lm$
incandescent standard	×			×	×	×		×		×	×	
bazat pe led-uri		×						×		×	×	
bec fluorescent			×					×		×		×
bec cu halogen		×		×			×		×		×	

Tabela 1.3: Același context ca în tabelul 1.2, dar transformat în context cu valori simple

Algoritmii descriși se împart în două categorii

- algoritmi de construcție a laticei de concepte a unui context
- algoritmi de afișare eficientă a laticei de concepte

1.3.1 Algoritmi pentru construirea laticei de concepte

Un algoritm simplu de înțeles pentru construirea laticei de concepte a unui context este cunoscut sub numele de *Next Neighbors* - **Vecinii următori**. Numele explică ideea de bază a algoritmului: generarea iterativă a vecinilor (elementelor *acoperite*), începând dintr-un concept, în relație cu $\prec$.

Conform [CR04], algoritmul este analog unei parcurgeri în lățime a laticei finale. O variantă bazată pe parcurgea în adâncime e de asemenea posibilă.

Începând cu elementul din vârful laticii (G, G'), algoritmul construiește un nivel odată, unde următorul nivel conține conceptele acoperite de toate conceptele prezente în nivelul curent. Mai exact, pentru fiecare concept din nivelul curent o funcție *GasesteElementeleAcoperite* (descrisă mai jos 1.3) e apelată care calculează conceptele acoperite de acel concept; funcția se bazează pe observația că toate conceptele acoperite de un concept se află într-o submulțime redusă de concepte - fiecare se obține prin a adăuga un atribut nou conținutului ($Y_1 = Y \cup \{m\}$) și a calcula (Y_1', Y_1'') apoi se verifică dacă fiecare concept acoperit rezultat nu a mai fost generat anterior, caz în care conceptul este adăugat laticei și în final conceptul e legat de părintele său (e adăugată muchia).

Laticea rezultată este strucutrată ca un tuplu de două mulțimi, (C, E), C fiind mulțimea conceptelor rezultate și E mulțimea laturilor dintre acestea, iar laturile sunt reprezentate ca și perechi ordonate de concepte $(c_1, c_2), c_1, c_2 \in C, c_1 < c_2$.

Date de intrare: Context (G, M, I)
Date de ieșire: Latice (C, E)

funcție VECINII URMĂTORI(Context (G, M, I))
    $C := \{(G, G')\}$
    $E := \varnothing$
    $nivelulCurent := \{(G, G')\}$
    cât timp $nivelulCurent \neq \varnothing$ **execută**
        $nivelulUrmator := \varnothing$
        pentru $(X, Y) \in nivelulCurent$ **execută**
            $elementeleAcoperite := GasesteElementeleAcoperite(X, Y)$
            pentru $(X_1, Y_1) \in elementeleAcoperite$ **execută**
                dacă $(X_1, Y_1) \notin C$ **then**
                    $C := C \cup \{(X_1, Y_1)\}$
                    $nivelulUrmator := nivelulUrmator \cup \{(X_1, Y_1)\}$
                $E := E \cup (X, Y) \to (X_1, Y_1)$
        $nivelulCurent := nivelulUrmator$
    întoarce $L = (C, E)$
funcție $GasesteElementeleAcoperite((X, Y))$
    $candidati := \varnothing$
    pentru $m \in M \backslash Y$ **execută**
        $X_1 := (Y \cup \{m\})'$
        $Y_1 = X_1'$
        dacă $X_1, Y_1 \notin candidati$ **then**
            $candidati := candidati \cup \{(X_1, Y_1)\}$
    întoarce $candidati$

Figura 1.3: Algoritmul Vecinilor următori, sau a Elementelor acoperite Sursa [CR04]

Avantajul acestui algoritm față de alții care au același scop (de a determina laticea de concepte a unui context) este că generează atât mulțimea conceptelor a unui context dar și diagrama (*conexiunile*) acestuia.

Complexitatea algoritmului este de $O(C \times G \times M^2)$. Deși există algoritmi cu o complexitate mai mică (neglijabil, conform [CR04]), algoritmul prezentat a fost ales deoarece este ușor de urmărit și prezintă conceptele descrise în prima secțiune.

1.3.2 Algoritmi pentru vizualizarea laticei de concepte

De-a lungul acestei lucrări, vorbim foarte mult despre diagramele laticelor de concepte și despre importanța *explorării* acestora, dar până acuma nu am menționat cum sunt generate aceste diagrame. Vizualizarea diagramelor este un domeniu dificil, întrucât un algoritm de generare trebuie să fie, în cele din urmă optimizat pentru lizibilitatea diagramei pentru oameni. Dispunerea nodurilor în așa fel încât să existe cât mai puține intersecții (acestea scad lizibilitatea diagramei drastic), păstrarea unei diagramei între anumite dimensiuni care să permită

înțelegerea conceptelor sunt doar câteva dintre provocările pe care le prezintă acest domeniu. Viteza de generare a diagramei este un alt criteriu important, mai ales când avem de-a face cu contexte complexe.

Dincolo de orice încercări de a reduce dimensiunile diagramelor, în cazul datelor reale, cu sute, sau chiar mii de concepte, majoritatea diagramelor laticelor de concepte se află în această situație. Acest lucru relevă nevoia pentru navigatoare dinamice a contextelor. Acestea trebuie să poată ascunde sau afișa doar părți dintr-o diagramă, în funcție de cerințele utilizatorului și a manipulărilor pe care acesta le realizează asupra conceptelor, relațiilor dintre ele, etc.

Există trei metode folosite pentru a comprima aceste diagrame la dimensiuni care să le facă comprehensibile.

1.3.2.1 Directoare ierarhice

Grafurile complexe sunt în general greu de vizualizat, așa că o metodă de a evita problemele asociate cu grafurile, putem reprezenta laticea ca un simplu arbore de ierarhii.

Luând în considerare relațiile de incluziune dintre concepte, ne este ușor să imaginăm laticea ca un arbore; elementul 1 al laticei devine rădăcina. Problemele cu acest mod de reprezentare intervin în momentul în care întâlnim două căi către același concept - un concept poate avea mai mulți *părinți*.

Cu toate acestea, pentru aplicații foarte simple, această metodă de vizualizare are meritele ei, în special când luăm în considerare obișnuința oamenilor cu organizarea arborescentă a datelor, datorită familiarității cu sistemele de fișiere și alte organizări ierarhice reprezentate ca arbori.

1.3.2.2 Diagrame imbricate

Unul din principiile de bază a vizualizării datelor este de a nu-ți copleși utilizatorul cu date. Asta se-ntâmplă când cineva deschide un fișier Excel cu zeci de rânduri și sute de coloane, sau când urmărește indicatorii bursei. De multe ori, pentru a evita acest fenomen, se caută separarea vizualizărilor în bucăți semi-independente, care pot fi înțelese.

În cazul diagramelor conceptelor, acest lucru se poate realiza prin **imbricarea** unor diagrame în diagrama principală.

De regulă, procedura se desfășoară în 4 pași ([CR04]):

1. Se partiționează mulțimea atributelor care descriu un context în două mulțimi.
2. Se găsesc laticile L_1 și L_2 corespondente subcontextelor descriise cu ajutorul mulțimilor definite în pasul anterior.
3. Se copiază L_2 în fiecare nod al diagramei lui L_1
4. Se diferențiază nodurile din fiecare copie a L_2 (de obicei prin colorare) care aparțin și de laticea completă.

L_1 devine astfel un cadru, schelet în care e imbricată apoi L_2. Matematic vorbind, laticea completă a conceptelor e conținută în produsul direct al laticelor de subcontexte ca o inf sublatice.

Diagramele imbricate sunt cele mai valoroase când expunem contexte cu valori multiple; acestea ajută la descoperirea unor subcontexte relevante.

Această metodă de simplificare a diagramelor este folosită de ToscanaJ, după cum se poate observa din figura 2.2.

Problemele acestei metode devin aparente în momentul în care încercăm să imbricăm o latice mai complexă. Indiferent de dimensiunea afișajului disponibil, fără un sistem de zoom perfecționat, de la un anumit număr de concepte ale subcontextului, laticea devine ilizibilă, pierzându-și orice valoare ca metodă de vizualizare. Astfel, în momentul organizării datelor trebuie avut grijă ca subcontextele care vor fi vizualizate să fie destul de mici încât să poată releva informații.

1.3.2.3 Vizualizări bazate pe focalizare+context

Prin această tehnică se dă o proeminență mai mare unei anumite zone a vizualizării. Se pornește de la premiza că utilizatorul caută ceva în latice, și atunci vizualizarea se comportă ca o lupă, mărind zona care, în funcție de anumite date de intrare consideră căar fi de interes. De obicei, această *zonă* este defapt unul din conceptele afișate.

Spre deosebire de diagramele imbricate, unde numai o submulțime a atributelor din context sunt afișate odată, tehnica aceasta folosește întreaga latice de concepte pentru a construi diagrama.

Marele avantaj al acestei metode de vizualizare este navigarea mult mai naturală față de celelalte metode, care restricționează accesul la foarte multe informații, ascunzându-le în diagrame secundare. În schimb, în această tehnică, informațiile sunt afișate cu grade diferite de detaliu, radiind dinspre punctul de focalizare al utilizatorului într-un moment oarecare (de unde și numele tehnicii).

Pentru a simula această *focalizare*, se vor modifica dimensiunile nodurilor, distanța față de nodurile adiacente, mărimea textului.

Determinarea punctului de interes se poate face în diferite feluri, urmărind cursurul mouse-ului, de exemplu. O altă posibilitate este de a-i permite utilizatorului să ofere direct diferite grade de interes unor noduri. Prin relațiile de vecinătate/acoperire, se pot calcula apoi gradele de interes pentru toate nodurile din latice, putând apoi să filtrăm totul ce scade sub un anumit grad de interes.

1.3.3 Filtrarea conceptelor după restricții date de utilizator

Cea mai eficientă metodă de a elimina din conceptele "inutile" este de a-i permite utiliztorului să restricționeze în mod repetat conceptele pe care le navighează, folosindu-se de cunoștiințele asupra domeniului, sau asupra scopului navigării.

Vom defini un cadru simplu pentru a permite utilizatorului să aplice dinamic restricții asupra conceptelor căutate și care vor face sistemul să elimine cele care nu se conformează restricțiilor. Folosind acest cadru, aplicând tot mai multe constrângeri, spațiul de căutare va converge spre țintă.

Vom defini constrângerile ca operatori definiți asupra unei sub-mulțimi de atribute. Fie laticea ne-filtrată $(C, \leq)$; asupra sa vom defini patru tipuri de constrângeri:

1. mulțimea elementelor sus c_1, adică $\uparrow c_1 := \{c \in C | c_1 \leq c\}$,

2. mulțimea elementelor jos c_1, adică $\downarrow c_1 := \{c \in C | c \leq c_1\}$

3. complementul lui sus c_1, adică $\neg \uparrow c_1 := \{c \in C | c_1 \neg \leq c\}$,

4. complementul lui jos c_1, adică $\neg \downarrow c_1 := \{c \in C | c \neg \in c_1\}$

Aceste constrângeri pot fi foarte ușor vizualizate grafic, raportat la partițiile impuse asupra spațiului de căutare. Se va observa că spațiul de căutare admis de aceste restricții sunt sub-latice de-ale $(C, \leq)$.

Definiție 16. *Fie două mulțimi S, G. Ele se numesc **mulțimile limită** ale spațiului de căutare, unde S conține mulțimea celor mai specifice concepte consistente (dacă un concept este **consistent** înseamnă că respectă toate restricțiile impuse), iar G conține mulțimea celor mai generale concepte consistente. Un concept c este admisibil dacă $\exists s \in$ astfel încât $s \leq c$ și $\exists g \in G$ astfel încât $c \leq g$.*

Constrângerile de tipul $\uparrow, \neg \downarrow$ mută S "în sus", în timp ce constrângerile $\downarrow, \neg \uparrow$ mută G "în jos". Când (sau dacă) cele două mulțimi sunt egale, ele vor conține doar conceptele *țintă*.

Constrângerile se pot recalcula iterativ, fără a recalcula spațiul eliminat de la început de fiecare dată când o constrângere este adăugată. Conform [CR04]:

$\uparrow c_1$:

$$G_{k+1} = \{g \in G_k | c_1 \leq g\}$$
$$S_{k+1} = min\{c \in C | c_1 \leq c, \exists s \in S_k : s \leq c, \exists g \in G_{k+1} : c \leq g\}$$

$\downarrow c_1$:

$$S_{k+1} = \{s \in S_k | s \leq c_1\}$$
$$G_{k+1} = max\{c \in C | c \leq c_1, \exists g \in G_k : c \leq g, \exists s \in S_{k+1} : s \leq c\}$$

$\neg \uparrow c_1$:

$$S_{k+1} = \{s \in S_k | c_1 \neg \leq s\}$$
$$S_{k+1} = max\{c \in C | c_1 \neg \leq c, \exists g \in G_k : c \leq g, \exists s \in S_{k+1} : s \leq c\}$$

$\neg \downarrow c_1$:

$$S_{k+1} = \{s \in S_k | s \leq c_1\}$$
$$G_{k+1} = max\{c \in C | c \leq c_1, \exists g \in G_k : c \leq g, \exists s \in S_{k+1} : s \leq c\}$$

Figura 1.4: Deducerea noilor mulțimi G și S din mulțimile curent G și S, pentru fiecare tip de constrângere

1.4 Utilizări practice

În cele 3 decenii de existență, analiza conceptuală formală și-a dovedit valoarea practică în mod repetat, în domenii foarte diverse.

Un studiu în domeniul cancerului s-a folosit de analiza conceptuală formală pentru a identifica bio-markeri(adică mulțimi de gene a căror schimbare în manifestare e corelată strâns cu boala), care pot să facă distincția între ținut în metastază și tumoare.[MVS08]

Forțele de ordine din lumea întreagă folosesc analiza conceptuală formală pentur a clasifica amenințări[VJ], a detecta comportamentul premergător extremiștilor musulmani [EPV+10] sau pentru a clasifica rapoarte de violență în violență domestică sau non-domestică.

Un meta-studiu asupra lucrărilor legate de analiza conceptuală formală a scos la lumină peste 1000 de articole științifice care menționează folosirea acestor tehnici în diferite domenii [PIKD13]. Din abstractul lucrării:

> "oferim [...] o vizune de ansamblu extensivă asupra publicațiilor dintre 2003 și 2011 în care au fost folosite metode bazate pe analiza conceptuală formală pentru acumulare de cunoștiințe și construcție de ontologii în diferite domenii. Aceste domenii includ minarea software-ului, analiza traficului web, medicină, biologie și chimie."

Capitolul 2
Situația actuală a software-ului existent în domeniu

În clipa de față există multe programe folosite pentru diferite aspecte ale analizei conceptuale formale. O listă mai dezvoltată, care adună majoritatea programelor disponibile poate fi găsită la [Pri07].

Mai jos vom discuta doar câteva programe care au influențat dezvoltarea Romagnei, sau sunt relevante din motive istorice, arhitecturale, etc.

2.1 Navigatoare de concepte

2.1.1 Toscana

Toscana[VW95] a fost lansat în 1995, și a fost unul din cele mai folosite unelte de explorare a laticelor de concepte pe parcursul următorilor ani. Aplicația a fost scrisă în C și folosea un format de date proprietar, bazat pe text.

Toscana este astăzi probabil cel mai bine ținută minte ca precursorul programului ToscanaJ, folosit și astăzi, prezentat în detaliu mai jos (2.1.2).

În ciuda eforturilor noastre, unealta nu a fost găsită pentru descărcare.

2.1.2 Toscanaj

ToscanaJ([Dev14b]) este "moștenitorul direct" al lui Toscana, lucru evidențiat și de nume (J-ul vine de la Java).

După cum spune chiar site-ul programului [Dev14c]: "E o unealtă de vizualizare pentru scheme conceptuale foarte avansată, care reușește să afișeze informație interogată dintr-o bază de date în diagrame de latice, sau direct din structuri de date luate din memorie."

2.1.2.1 Funcționalități

Din nou, citând site-ul programului[Dev14c], prezentăm câteva funcționalități ale programului:

"

- Afișarea diagramelor simple și imbricate.
- Culoarea unui nod reprezintă mărimea contingentului obiectelor (poate fi modificat să reprezinte cuprinsul), deasemenea mărimea nodului poate fi folosită pentru același tip de informație.
- Mulțimea de obiecte de interes poate fi filtrată printr-un dublu click asupra nodurilor

- Nodurile din diagramă pot fi selectate pentru a fi scoase în evidență, pentru a ajuta citirea *[n.t. diagramei]*.
- Diagramele pot fi exportate ca SVG, PNG și JPEG. Informații adiționale despre cum diagrama a fost obținută sunt exportate ca fișiere text separate, prin memoria temporară a calculatorului *(clipboard)*, sau direct în fișierul SVG (ca elementul `<desc>`).
- Etichetele nodurilor pot avea conținut diferit, folosindu-se de fragmentele de SQL specifice datelor
- Vizualizări adiționale a bazei de date pot fi deschise din diagramă, de exemplu folosind șabloane HTML în care rezultatele interogărilor sunt afișate.
- Interfața de vizualizare a bazei de date a fost gândită ca o interfață pentru pluginuri pentru a ușura extinderea ToscanaJ pentru scopuri specifice.
- Descrieri HTML pot fi atașate schemei, diagramelor și atributelor
- Vederile bazelor de date pot fi folosite pentru atribute, de exemplu pentru a interoga un URL din baza de date care e mai apoi deschis într-un navigator extern.

”

Având în vedere că scopul programului Romagna este de a oferi o alternativă programului Toscana(J), aceste funcționalități se vor regăsi și în Romagna, alături de altele, descrise în capitolul 4.1.

Mai jos prezentăm două capturi de ecran care prezintă programul ToscanaJ.

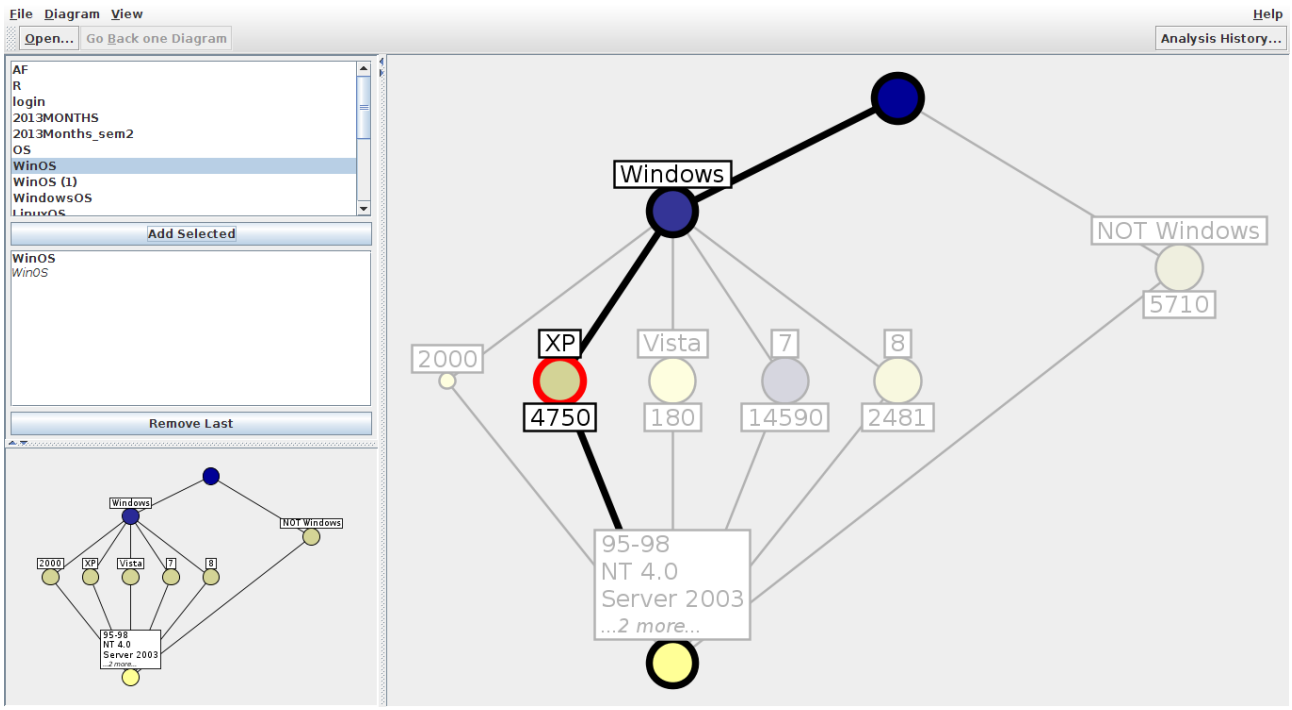


Figura 2.1: ToscanaJ, afișând o latice simplă generată din date de acces ale unui site web

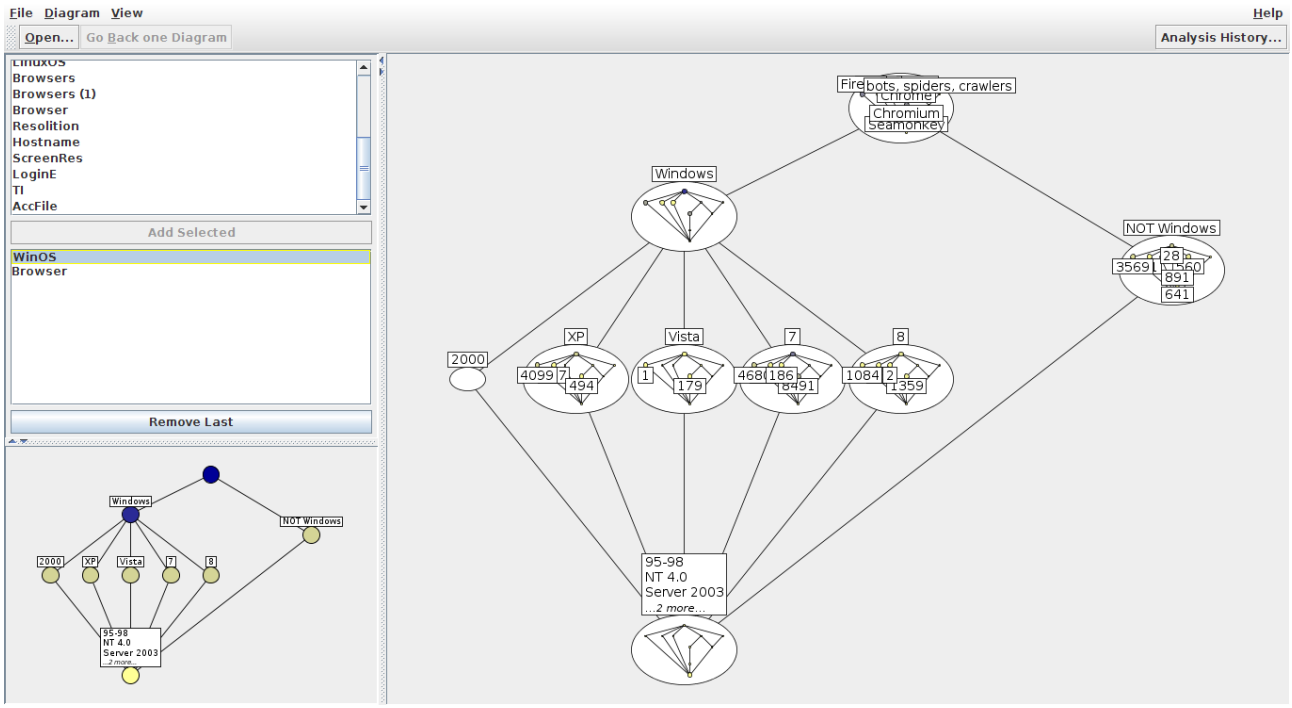


Figura 2.2: ToscanaJ, afișând aceeași latice ca în 2.1, de data aceasta imbricată cu altă diagramă

Interfața ToscanaJ este punctul de pornire pentru Romagna. Vom încerca să păstrăm anumite elemente comune, pentru a ajuta utilizatorii obișnuiți, dar vom face, bine-nțeles schimbări și adaptări, mai ales luând în considerare diferențele convențiilor dintre interfețele aplicațiilor web, și cele desktop.

ToscanaJ este, asemenea Romagna, doar un navigator de concepte. Modelarea conceptelor din date este realizată cu ajutorul altor programe din suita din care face parte și ToscanaJ.

Elba este un editor pentru schemele conceptuale, legat de baze de date.

Siena este un editor pentru scheme conceptuale, foarte asemănător cu **Elba**, diferența fiind că Siena nu are nevoie de o legătură cu o bază de date, putând descrie concepte simple direct în program.

2.1.3 GaloisExplorer

GaloisExplorer [tea09] este un program mai recent, cu ultima actualizare în 2009.

Are o intefață realizată în QT, ceea ce îi permite să funcționeze pe mai multe platforme. O abordare interesantă, dar în cele din urmă doar cu valoare estetică, este prezentarea laticelor în 3d (2.3).

Din păcate, folosește anumite librării 3d (coin3d[**?**]) care nu sunt imediat disponibile și care necesită compilare manuală.

Pentru mulți utilizatori care nu au cunoștințe avansate de calculatore, aceste impedimente sunt motiv suficient pentru a abandona această instalare.

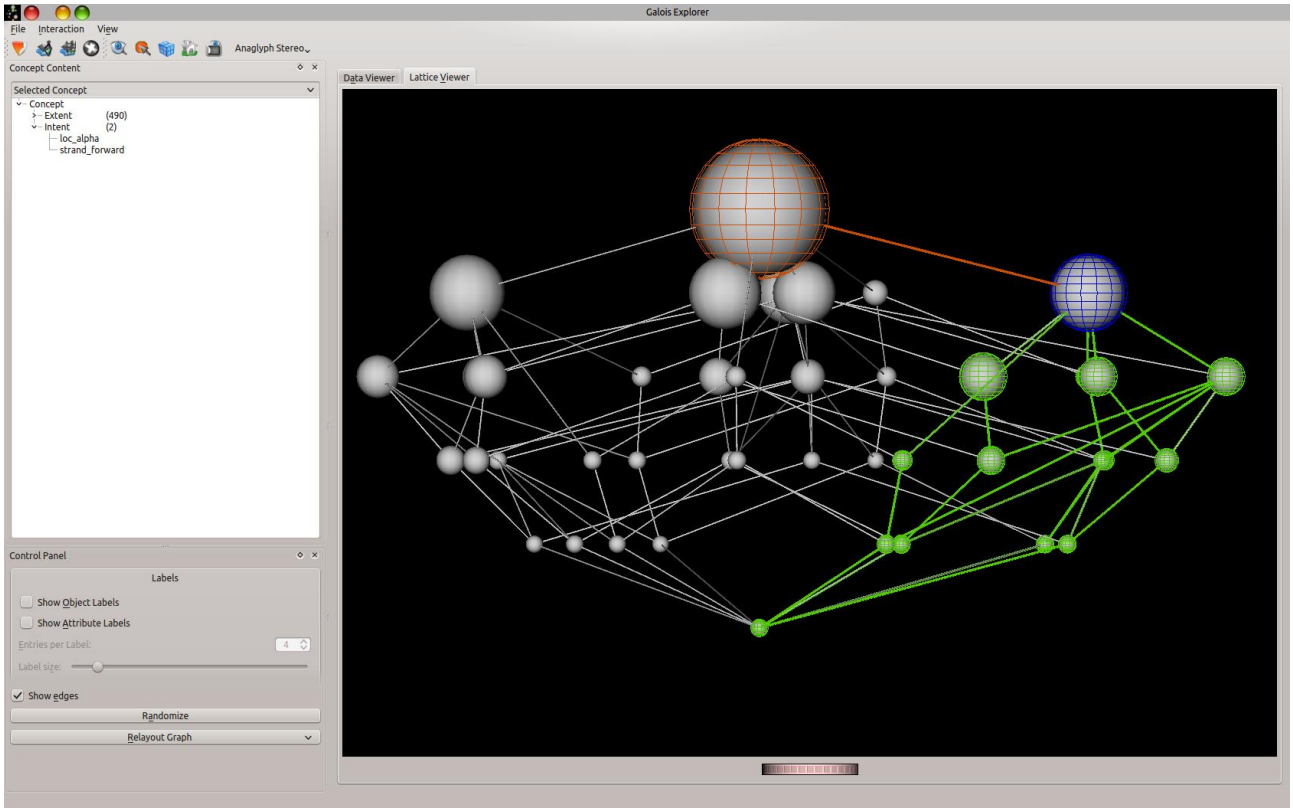


Figura 2.3: GaloisExplorer, afișând o diagramă în 3d. Sursa: site-ul proiectului [tea14]

2.1.4 Galicia

Galicia este o aplicație Java dezvoltată la Universitatea din Montreal care dorește să acopere toate funcționalitățile necesare lucrului cu latice și contexte. Permite atât construcția și editarea contextelor, cât și vizualizarea 1aticilor rezultate.

2.2 Software conex

În această secțiune vom prezenta câteva alte programe remarcabile în domeniul FCA, sau care au avut un impact indirect asupra dezvoltării Romagna:

FCAStone dezvoltat de Uta Priss, e un program care urmărește obținerea inter-operabilității între diferite programe pentru FCA (numele vine de la Piatra *[n.t. Stone = Piatră]* Rosetta). Astfel, convertește fișierele folosite de mai multe suite în FCA. Pe lângă asta, convertește contexte în latice de concepte, și latice de concepte în formate grafice (atât vectoriale cât și raster).

Conexp dezvoltat de Serhiy A. Yevtushenko, este o unealtă scrisă în Java, folosită în principal în scopuri educaționale (dar nu numai), are numeroase funcționalități de modelare a contextelor, dintre care amintim (preluate de site-ul programului [Yev14]):

- calculează numărul de concepte formale dintr-un context dat
- calculează laticea de concepte
- permite explorarea atributelor

- calculează regulile de asociații

conexp-clj O reimplementare scrisă de Daniel Borchmann în limbajul de programare Clojure a programului de mai sus. Programul pune accentul pe linia de comandă și programarea interactivă în domeniul FCA. Recent a adăugat un modul de interfață grafică, încă în stadiul experimental.

OpenFCA Este o suită de trei aplicații, din care Conflexplore este o aplicație bazată pe Flex, folosită pentru explorarea conceptelor, atât tabelar cât și printr-o formă simplă de latice de concepte.

Lattice Miner Este o aplicație scrisă în Java, bazată pe 3 module, pentru concepte, contexte și reguli de asociere. Având structură modulară, permite integrarea de plug-inuri. Suportă interoperabilitatea cu Galitia și ConExp.

LaTeXfor FCA [Gan14] Plugin pentru LaTeX, scris de Bernhard Ganter, oferă câteva "environment"-uri noi, folosite și în această lucrare, care permite aranjarea ușoară în pagina a contextelor și a laticilor derivate din acestea.

Capitolul 3
Tehnologii folosite

După cum am menționat, Romagna se bazează pe tehnologii web. Vom continua prin a descrie sumar principalele librării/framework-uri folosite și un raționament scurt pentru existența lor în proiect.

3.1 Extinderea trăsăturilor de limbaj - CoffeeScript

Unul din principalele motive pentru care browser-ul nu a fost mult timp acceptat ca o platformă matură pentru aplicații era absența altor limbaje de programare disponibile înafară de JavaScript.

Această stare de fapt se schimbă încet din mai multe motive:

- JavaScript a evoluat și este considerat un limbaj de programare mai matur decât în perioada în care a fost scris ToscanaJ.
- Existența **emscripten**, descris pe scurt în subsecțiunea 3.4.1 și rezultatele obținute de acesta.
- Existența limbajelor care compilează *în* JavaScript.

Unul din aceste limbaje este **CoffeeScript** [A+12]. CoffeeScript nu introduce decât două concepte (clase versus prototipuri și array comprehensions) noi față de JavaScript, dar simplifică dezvoltarea prin câteva îmbunătățiri:

- Evitarea variabilelor globale
- Câteva scurtături de sintaxă (`->` pentru a declara o funcție, `=>` pentru a declara o funcție legată de contextul actual, etc.)
- Clarificarea operatorilor de comparație (`==` în CoffeeScript devine automat `===` în JavaScript, scăpând automat de o întreagă clase de erori).

Alături de sintaxa mai prietenoasă (o opinie personală a autorului), prin elidarea acoladelor în favoarea indentării pentru definirea blocurilor, am ales CoffeeScript ca limbajul principal pentru Romagna.

3.2 Scalabilitate și arhitectură corectă în aplicații web - Ember.js

Ember.js este un framework scris în JavaScript, inspirat de Cocoa, și are ca scop ușurarea dezvoltării aplicațiilor web complexe, prin oferirea unui cadru MVC, cu o structură relativ rigidă.

Inițial aplicația nu era bazată pe niciun framework. Pe măsură ce complexitatea structurii a crescut, am realizat că recreem, involuntar, un sistem MVC, mai mult ca sigur imperfect.

Am hotărât să reorganizăm codul în urma unei deliberări, bazându-ne pe faptul că un framework folosit de mii de oameni, cu teste funcționale și modulare la zi poate oferi o fundație puternică, fără a ne cheltui timpul "re-inventând roata". Din variantele disponibile de framework-uri pentru JavaScript, am ales Ember.js deoarece:

- Structura impusă se potrivește cu necesitățile proiectului.
- Posibilitatea schimbării ușoare a adaptorului a repozitoriului de date inclus în framework înseamnă că vom putea crea o variantă viitoare a Romagna care se includă comunicarea cu un server, funcționalitate care este plănuită pentru versiuni care vor urma.
- Funcționalitatea componentei `Router`, care permite crearea de URL-uri, care pot fi salvate de utilizator, în funcție de navigarea sa prin aplicație, înseamnă că avem la îndemână o formă simplă de serializare și salvare a stării aplicației.

3.3 Manipularea documentelor și vizualizare de date - d3.js

D3 (provenit din *Data-Driven Documents* - Documente Bazate pe Date) este o librărie scrisă în JavaScript pentru manipularea DOM-ului relaționând cu date legate de elemente din document. Citând din abstractul lucrării [BOH11]:

> "Cu D3, designerii leagă în mod selectiv date de intrare de elemente arbitrare ale documentului, aplicând transformări dinamice pentru a genera, cât și a modifica conținut."

Devine clar din descrierea de mai sus că această librărie va ușura dezvoltarea aplicației, datorită posibilității de a lega *conceptele* de reprezentarea acestora într-un mod care să permită modificarea acestei reprezentări în funcție de explorarea diagramelor de către utilizator, ceea ce, este, până la urmă scopul aplicației.

3.3.1 Grafică scalabilă - svg

D3 se bazează pe manipularea DOM-ului pentru a obține vizualizări dinamice de date. Deși HTML-ul poate fi stilizat, formele complexe și imaginile vectoriale sunt descrise în browser prin SVG [Fer01], un standard de grafică vectorială vazată pe XML. Deși ajută la crearea unor aplicații de vizualizare complexe, SVG-ul vine cu problemele sale, care sunt descrise mai pe larg în sub-subsecțiunea 4.4.1.2.

3.4 SQL în browser - sql.js

După cum am descris în secțiunea 4.1.2, pilonii pe care am decis să construim Romagna (ușurința utilizării și păstrarea locală a datelor) pun în dificultate folosirea unei arhitecturi client-server, deoarece vrem un proces de instalare și folosire cât mai simplu (nici un proces de instalare nu e mai simplu decât deschiderea unei pagini web), iar utilizatorii nu-și vor trimite

datele unei părți terțe, din temeri proprii sau pentru că nu au acest drept, lucrând cu date confidențiale.

Astfel, compromisul este de a analiza SQL în browser, cu ajutorul sql.js [Zak14], care este defapt SQLite re-compilat în JavaScript cu ajutorul emscripten.

SQL.js poate rula într-un Web Worker, echivalentul unui thread izolat folosit de JavaScript în browser [Hic12]. Astfel, operațiile de interogare asupra bazei de date nu au un impact direct asupra răspunsului interfeței.

3.4.1 Compilarea codului C/C++ în JavaScript - emscripten

Din abstractul articolului [Zak11]

> "[...] prezentăm Emscripten, un compilator din limbaj de asamblare LLVM (Low Level Virtual Machine) în Javascript. Acesta deschide două căi de a rula cod scris în alte limbaje decât Javascript pe web: 1. Compilarea codului direct în limbaj de asamblare LLVM și apoi compilarea acestuia în JavaScript folosindu-vă de Emscripten, sau 2. compilarea întregului run-time al unui limbaj interpretat în LLVM și apoin JavaScript."

Impactul pe care acest compilator l-a avut este, cel puțin până acum, unul mai mult teoretic, majoritatea programelor fiind folosite cu rol demonstrativ. Aplicațiile compilate în JavaScript sunt, binențeles, mai lente decât variantele lor compilate direct în limbaj de asamblare. Totuși pentru o anumită categorie de programe, între care se află și Romagna, acest schimb de viteză pentru funcționalitățile oferite de librăria sau aplicația compilată este binevenit. Structurile de date folosite de Romagna până în acest moment nu sunt atât de mari încât pierderea performanței să fie critică.

Capitolul 4
Romagna

Romagna este o aplicație dezvoltată începând cu anul 2014, pornită la Universitatea Babeș-Bolyai ca o alternativă folosindu-se de tehnologiile web pentru un navigator de concepte modern.

Numele (Romagna) este o referință directă suitei de programe care a inspirat această aplicație, Toscana. Romagna este o regiune istorică a Italiei aflată la nordul Toscanei.

Pe lângă funcționalitățile ale aplicației ToscanaJ descrise în sub-subsecțiunea 2.1.2.1, Romagna are câteva funcționalități noi plănuite, descrise în secțiunea 4.4.1.2.

Dezvoltarea acestei aplicații are câteva scopuri *(diferite de funcționalități)* principale, prezentate în secțiunea următoare.

4.1 Raționament

4.1.1 Ușurința utilizării

Uneltele destinate FCA trebuie gândite ca programe pentru consumatori, utilizatori obișnuiți. Publicul țintă al acestor aplicații nu sunt neapărat programatori sau oameni cu experiență în folosirea avansată a calculatoarelor. Oamenii care beneficiază cel mai mult de asemenea aplicații sunt cercetători în domenii foarte diferite (după cum se vede în secțiunea de aplicații practice 1.4)

Din acest punct de vedere, este părerea autorului că soft-ul existent în acest domeniu nu satisface necesitățile publicului său țintă.

Pentru clarificare, următoarele puncte trebuie îndeplinite pentru a considera o aplicație ușor de folosit:

- Aplicația trebuie să fie **ușor de instalat**. Programatorii de multe ori scapă din vedere acest aspect, dacă nu e vorba de un produs comercial, care are nevoie de cât mai mulți utilizatori. E inadmisibil să se ceară compilarea manuală a unei librării C++, sau instalarea unui server MySQL unui utilizator obișnuit pentru a utiliza un program. Romagna, fiind o aplicație web, sare peste acești pași. Singurii pași pre-mergători folosirii aplicației sunt pregătirea datelor pentru consum de către aplicație și familiarizarea cu interfața aplicației.

- Aplicația trebuie să fie **disponibilă pe mai multe platforme**. Utilizatorii nu vor instala un alt sistem de operare pentru a testa o aplicație. În schimb, majoritatea utilizatorilor au la dispoziție un navigator web modern, cum ar fi Mozilla Firefox, Google Chrome, etc. Aceste medii de aplicații (pentru că navigatoarele web și mediul de execuție al limbajului JavaScript pus la dispoziție de acestea a devenit un mediu de aplicații) se actualizează

automat, permițând utilizarea celor mai noi funcționalități unei majorități a utilizatorilor acestora.

- Aplicația trebuie să ofere **date de ieșire utile** utilizatorului. Ne referim aici bine-nțeles la rezultatele așteptate de utilizator, prezentate într-un mod clar, dar și la mesajele de erori, care trebuie să fie inteligibile, să ofere cursuri de acțiune care să rezolve problema întâmpinată, într-un limbaj non-tehnic, pe cât posibil. Erorile criptice, cum ar fi stack-trace-uri, sau excepții afișate pe ecran descurajează utilizatorul obișnuit.

- Aplicația trebuie să aibă **o interfață plăcută**. E părerea autorului că, deși utilizabile, interfețele generate de aplicații Java în general oferă o experiență neplăcută, nu se integrează în stilul sistemului de operare. Aplicațiile web sunt mult mai ușor de estetizat, mai ales cu ajutorul unor framework-uri CSS, dintre care amintim Twitter Bootsrap sau Zurb Foundation.

- Ca o ultimă alternativă pentru utilizator, aplicația trebuie să vină **însoțită de documentație**. Aceasta trebuie formulată într-un limbaj non-tehnic.

- Opțional, aplicația ar trebui să aibă **interfața localizată**. Nu toți utilizatorii de calculatoare sunt vorbitori de engleză. Folosirea pictogramelor pentru realizarea interfeței poate ajuta, dar inclusiv în acest caz, diferențele culturale, sau conceptele prea complexe pe care dezvoltatorii încearcă să le transmită, pot descuraja utilizarea aplicației mai mult decât să ajute.

Romagna îndeplinește unele din condițiile de mai sus(ușor de instalat, ușor de găsit) prin natura aplicațiilor web, iar celelalte condiții constituie obiectul efortului autorului.

4.1.2 Folosirea web-ului, păstrarea confidențialității datelor

Este o stare de fapt că în domeniul analizei datelor, integritatea și confidențialitatea acestora este o grijă constantă.

Mulți cercetători lucrează de asemenea cu date care sunt, legal vorbind, secrete. Se înțelege astfel, că majoritatea potențialilor utilizatori **nu** vor avea încredere să trimită datele unui serviciu extern.

Din aceste considerente, o arhitectură server peste web (și oferirea Romagnei ca serviciu) a fost imediat desconsiderată. Dar, cum am menționat în subsecțiunea 4.1.1, ușurința de utilizare este o altă calitate pe care încercăm să o regăsim în Romagna.

S-a ajuns astfel la compromisul (temporar) de a realiza **totul** în browser. Asta înseamnă accesarea datelor SQL în browser, prin tehnologii descrise în capitolul **??**.

4.2 Date de intrare

Datele de intrare pentru Romagna sunt (cel puțin în acest moment, pe viitor vom accepta mai multe tipuri de date) fișiere `.csx`, fișiere XML care au o structură aproximativă:

După cum se vede, fișierul e secționat în diagrame, care conțin conceptele ca și colecții de referințe către atribute și obiecte și muchii între concepte. Pentru brevitate, am emis elementele

```
<?xml version="1.0" encoding="UTF-8"?>
<conceptualSchema version="TJ1.0">
    <diagram title="Diagram Title">
        <node id="2">
            <position x="20.0" y="20.0" />
            <concept>
                <objectContingent />
                <attributeContingent>
                    <attribute>&gt;=2500$</attribute>
                </attributeContingent>
            </concept>
        </node>
        <node id="3">
            <position x="40.0" y="40.0" />
            <concept>
                <objectContingent />
                <attributeContingent>
                    <attribute>&gt;=3000$</attribute>
                </attributeContingent>
            </concept>
        </node>
        <edge from="2" to="3" />
        <edge from="2" to="9" />
    </diagram>
</conceptualSchema>
```

Figura 4.1: Exemplu de sintaxă a unui fișier de intrare.

¡objectElementStyle¿ și *¡attributeELementStyle*, care descriu cum ar trebui desenată eticheta fiecărui atribut/obiect în parte.

Dacă fișierul csx nu conține datele contextului, are nevoie de o bază de date. Aceasta este referențiată prin

```
<databaseConnection>
  <embed url="file:///home/calea/catre/baza_de_date.sql" />
    <table name="Tabel_relevant" />
    <key name="COOKIE_ID" />
</databaseConnection>
```

și referințele către obiecte/atribute se fac prin interogări care vor fi executate pe tabelul refernțiat mai sus.

```
<object>((IPID>8) AND (IPID<=9)) AND ((IPID>30) AND (IPID<=40))</object>
```

4.3 Structură

Structura la nivel înalt este ghidată de folosirea librăriei Ember.js. Aceasta impune dezvoltarea aplicației pe principiile MVC. Pe lângă aceste principii, cerințele specifice domeniului au dus la câteva deviații mici.

În mod normal accesul la date ar trebui să se facă *exclusiv* printr-un repozitoriu de date, dar datorită naturii fișierelor csx, care permit interogări sql integrate în structura lor (după

cum am arătat mai sus), controlorul are nevoie de acces direct la o bază de date, fără a accesa repozitoriul pus la dispoziție de Ember.

Întrucât există mai multe variante ale formatului `.csx` (nu este standardizat), primul pas este uniformizarea și prelucrarea datelor de intrare în structurile de date interne ale aplicației (care coincid în mare parte cu conceptele descrise în 1 - modele care descriu *contexte, concepte, atribute, obiecte, latice*). Pe lângă acestea, diagramele în sine sunt modele, și au controlori aferenți.

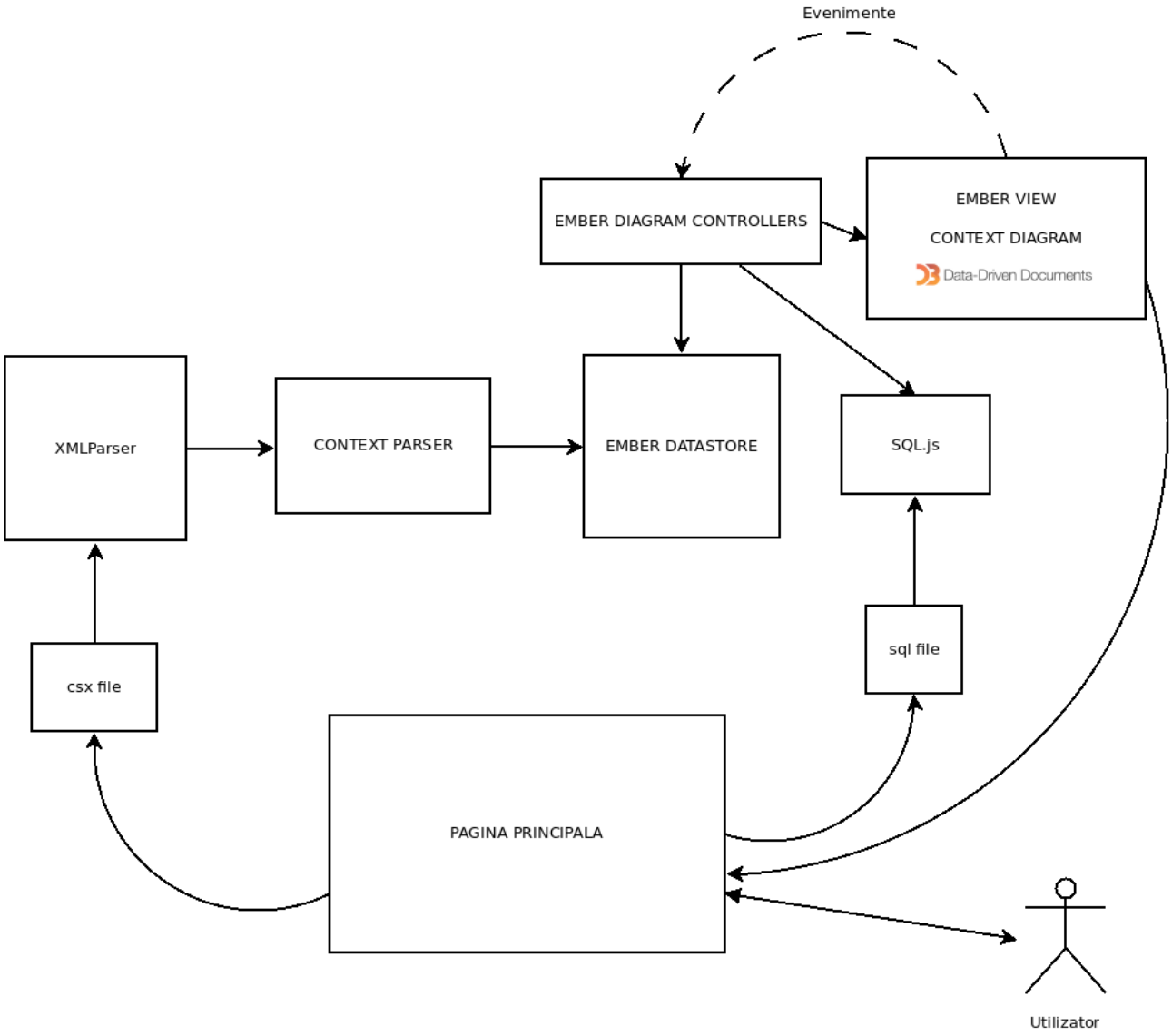


Figura 4.2: Diagrama viziunii de ansamblu asupra arhitecturii aplicației Romagna

View-ul, sau interfața diagramei este controlată de Ember doar în raport cu celelalte componente ale aplicației (mai specific cu controlorii, View-urile sunt, în MVC-ul clasic, pe care Ember îl respectă, limitate la a comunica cu controlorul); desenarea diagramei este făcută de către d3.js, iar tot el înregistrează evenimentele DOM, pe care le trimite mai departe la Ember, care, la rândul său, va interoga sql.js pentru date, sau își va interoga propriul repozitoriu de date, în cazul în care fișierul `.csx` conținea toate datele contextului.

4.4 Dezvoltare

Dezvoltarea a început în anul 2014. Odată alese tehnologiile pe care am dezvoltat aplicația, primiii pași au fost studiere în detaliu a fișierelor `.csx`, care sunt inputul principal al programului ToscanaJ.

În figura 4.3 se vede o versiune incipientă a programului afișând o diagramă extrasă dintr-un fișier de exemplu al ToscanaJ.

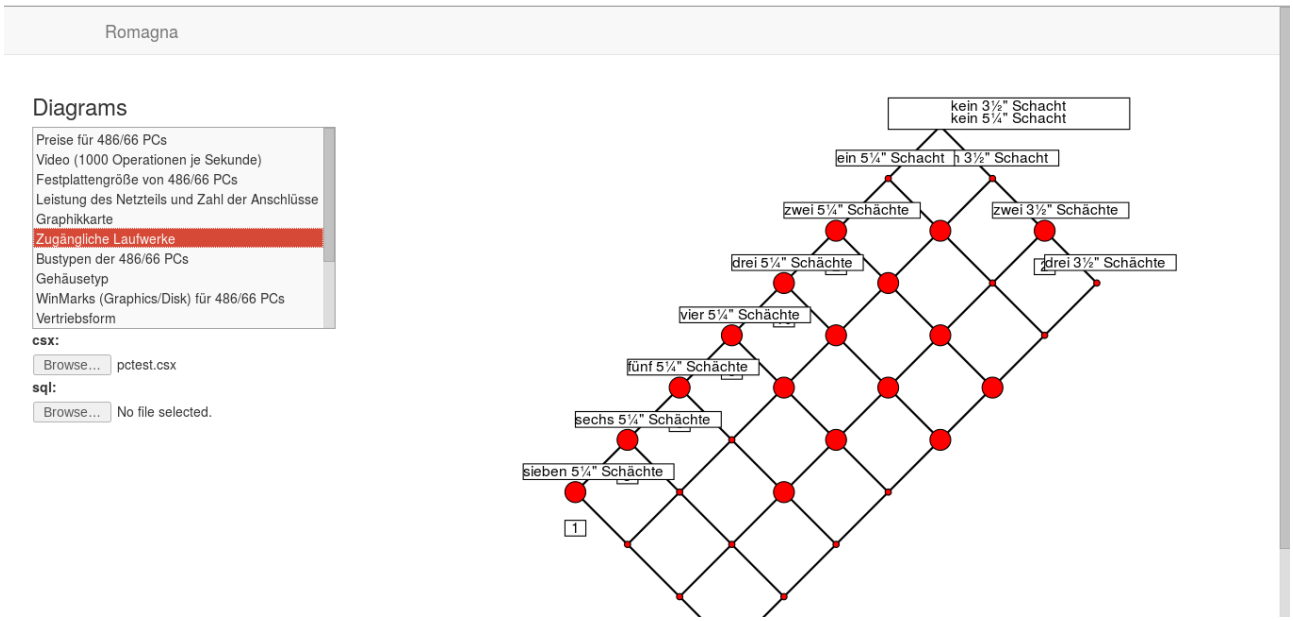


Figura 4.3: Versiunea 0.1 a aplicației, rulând pe Mozilla Firefox, versiunea 33, pe sistemul de operare Ubuntu Linux afișând o diagramă derivată dintr-un fișier de test inclus în arhiva descărcată cu ToscanaJ

4.4.1 Probleme întâmpinate

4.4.1.1 MySQL versus sql.js

ToscanaJ folosește baza de date MySQL pentru interpretarea volumelor mai mari de date decât ar fi gestionabile în fișiere excel sau csv, etc. În schimb, Romagna folosește o variantă a SQLite.

Din fericire, site-ul dezvoltatorilor SQLite oferă o listă [Dev14a] de programe folosite pentru convertirea fișierelor de export MySQL în fișiere de folosit pentru SQLite.

4.4.1.2 SVG și probleme în afișarea corectă a textului

HTML-ul (mai ales folosit în tandem cu CSS), la fel ca și LaTeX, de altfel, permite formatarea foarte ușoară a textului. Această proprietate derivă din scopul existenței sale ca un limbaj de formatare a documentelor inter-conectate.

SVG-ul în schimb are ca scop descrierea imaginilor vectoriale. Astfel, textul nu are un rol special în acest standard; toate uneltele cu care browser-ul este dotat pentru încadrarea textului în containere nu sunt disponibile atunci când lucrăm cu svg, cum este în cazul nostru la afișarea diagramelor.

Chiar dacă un element `<text>` se află într-un container cum ar fi un element `<rect>` sau `<g>` și containerul are dimensiuni bine definite, textul se va revărsa înafara containerului, ceea ce va duce la un aspect foarte urât și la probleme grave de lizibilitate, după cum se poate observa din figura 4.4

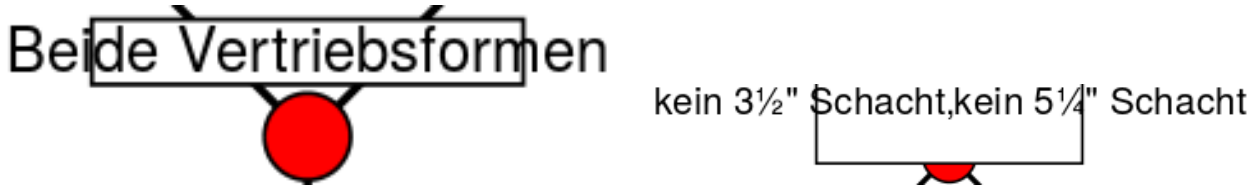


Figura 4.4: Captură de ecran din varianta 0.1 a aplicației, demonstrând text care se revarsă din containerul său.

Rezolvarea acestor probleme implică calcularea constantă a lungimii textului, redimensionarea containerului, sau în cazul în care containerele sunt aglomerate micșorarea textului.

De asemenea, începerea unui rând nou într-un element `<text>` nu este posibilă printr-o secvență de caractere (cum este `\n` în string-urile obișnuite, sau `\\` în LaTeX). Orice linie nouă trebuie introdusă într-un element nou `<tspan>`, și poziționat manual.

Astfel, multă muncă trebuie depusă în formatarea textului, un aspect care în celelalte părți ale dezvoltării web este rezolvat complet de platformă.

Concluzii

Dezvoltarea aplicației nu este un proces încheiat (și este părerea autorului că în majoritatea cazurilor procesul de dezvoltare al software-ului nu se sfârșește niciodată), dar putem ajunge la niște concluzii.

1. Analiza conceptuală formală este un câmp vast, din a cărui teorie am dezvoltat foarte puțin în această lucrare. Pentru mai multe detalii, recomandăm sursele citate în capitolul 1 și ale cărei aplicații concrete sunt prea numeroase pentru a fi numerotate.

2. Platforma web s-a dovedit a fi una potrivită pentru provocarea pe care o reprezintă navigarea conceptelor unui context.

3. Deși aplicația funcționează, în sensul de a avea *funcționalitățile* propuse, una din provocările pe care ni le-am propus când am început acest proiect a fost să creăm o aplicație care să fie plăcută de cercetători și de amatori de vizualizare. Întrucât proiectul nu a fost făcut public până acum, nu putem confirma că am reușit în această priviință.

Viitor

Dezvoltarea Romagna va continua în trei direcții mari în viitor:

Rafinarea interfeței existente Pentru primele versiuni a aplicației am pus accent pe o interfață cât mai simplă, de-a dreptul spartană. Vom continua prin a aduce îmbunătățiri cum ar fi funcționalitate pentru taburi, pentru a permite utilizatorilor explorarea mai multor subcontexte al aceluiași context simultan.

Alte îmbunătățiri planificate țin de posibilitatea explorării unei latice prin zoom și prin metode de focalizare+context, descrise în subsecțiunea 1.3.2.3.

Dezvoltarea unui editor live de contexte, cu răspuns imediat în afișarea diagramei O unealtă care, în opinia autorului, ar fi *foarte* utilă în scopuri didactice și exploratorii. A putea vedea direct implicațiile pe care schimbarea valorii unui atribut asupra unui obiect le are asupra contextului și laticei poate ajuta înțelegerea domeniului foarte mult de către studenți și amatori. Putem extinde acest concept după aceea la folosirea aplicației pentru dezvoltarea unei pagini complexe de prezentare a analizei conceptuale formale studenților, cu cursuri interactive direct în browser.

Această funcționalitate necesită însă un efort de dezvoltare foarte mare, deoarece pentru a prezenta aceste schimbări în timp real, fără a duce la frustrarea utilizatorului necesită algoritmi foarte eficienți de conversie a contextului în latice și de a reprezenta laticea (și schimbările acesteia).

Editarea contextului *din* diagramă În mod obișnuit diagrama are doar rol de prezentare. Dar putem să ne gândim la un caz de utilizare unde editarea contextului este posibilă direct din latice.

Dezvoltarea unui convertor automat din MySQL în SQLite - Momentan conversia datelor este un proces separat, pe care-l cerem utilizatorilor de ToscanaJ care doresc să încerce aplicația noastră. Acest pas în plus poate fi destul pentru anumiți utilizatori să renunțe să încerce aplicația noastră.

Dezvoltarea unei componente server Actuala configurație a întregii activități desfășurate în browser este un compromis, apărut în urma cerințelor. Un server dedicat găzduirii și procesării bazei de date va permite folosirea programului pentru seturi de date prea mari pentru a fi încărcate în browser.

Dezvoltarea în această direcție este dependentă de răspunsul din partea utilizatorilor.

Deși dezvoltarea a început ca un proiect de diplomă, autorul pune codul sursă la dispoziția oricui și încurajează colaborarea și/sau dezvoltarea în paralel pentru diferite necesități. Repozitoriul canonic se află pe site-ul de găzduire a software-ului Github [ea14], unde veți găsi informații curente despre dezvoltarea Romagnei.

Bibliografie

[A+12] Jeremy Ashkenas et al. Coffeescript. *Retrieved from* `http://coffeescript.org`, 2012.

[BOH11] M. Bostock, V. Ogievetsky, and J. Heer. D3; data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, Dec 2011.

[CR04] Claudio Carpineto and Giovanni Romano. *Concept Data Analysis: Theory and Applications*. John Wiley & Sons, 2004.

[Dev14a] SQLite Developers. Sqlite cvstrac. `http://www.sqlite.org/cvstrac/wiki?p=ConverterTools`, 2014. [Online; accesat la 18-Iunie-2014].

[Dev14b] ToscanaJ Developers. Toscanaj homepage. `http://toscanaj.sourceforge.net/`, 2014. [Online; accesat la 15-Iunie-2014].

[Dev14c] ToscanaJ Developers. Toscanaj homepage. `http://toscanaj.sourceforge.net/toscanaJ/index.html`, 2014. [Online; accesat la 15-Iunie-2014].

[ea14] Mihai Chereji et. al. romagna/romagna. `https://github.com/romagna/romagna/`, 2014. [Online; accesat la 19-Iunie-2014].

[EPV+10] Paul Elzinga, Jonas Poelmans, Stijn Viaene, Guido Dedene, and Shanti Morsing. Terrorist threat assessment with formal concept analysis. In *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*, pages 77–82. IEEE, 2010.

[Fer01] Jon Ferraiolo. Scalable vector graphics (SVG) 1.0 specification. W3C recommendation, W3C, September 2001. http://www.w3.org/TR/2001/REC-SVG-20010904.

[Gan14] Bernhard Ganter. Latex for fca. `http://www.math.tu-dresden.de/~ganter/fca/`, 2014. [Online; accesat la 15-Iunie-2014].

[GW97] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition, 1997.

[Hic12] Ian Hickson. Web workers. Candidate recommendation, W3C, may 2012. http://www.w3.org/TR/2012/CR-workers-20120501/.

[MVS08] Susanne Motameny, Beatrix Versmold, and Rita Schmutzler. Formal concept analysis for the identification of combinatorial biomarkers in breast cancer. In Raoul Medina and Sergei Obiedkov, editors, *Formal Concept Analysis*, volume 4933 of *Lecture Notes in Computer Science*, pages 229–240. Springer Berlin Heidelberg, 2008.

[PIKD13] Jonas Poelmans, Dmitry I. Ignatov, Sergei O. Kuznetsov, and Guido Dedene. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, 40(16):6538 – 6560, 2013.

[Pri07] Uta Priss. Formal concept analysis homepage. http://www.upriss.org.uk/fca/fcasoftware.html, 2007. [Online; accesat la 15-Iunie-2014].

[tea09] GaloisExplorer team. Galoisexplorer: Project web hosting - open source software. http://galoisexplorer.sourceforge.net/, 2009. [Online; accesat la 15-Iunie-2014].

[tea14] GaloisExplorer team. Galoisexplorer — free science & engineering software downloads at sourceforge.net. http://sourceforge.net/projects/galoisexplorer/, 2014. [Online; accesat la 15-Iunie-2014].

[VJ] Susan Voss and CA Joslyn. Advanced knowledge integration in assessing terrorist threats. Technical report.

[VW95] Frank Vogt and Rudolf Wille. Toscana — a graphical tool for analyzing and exploring data. In Roberto Tamassia and IoannisG. Tollis, editors, *Graph Drawing*, volume 894 of *Lecture Notes in Computer Science*, pages 226–233. Springer Berlin Heidelberg, 1995.

[Wil82] Rudolf Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In Ivan Rival, editor, *Ordered Sets*, pages 445–470, Dordrecht/Boston, 1982. Reidel.

[Yev14] Serhiy A. Yevtushenko. The concept explorer. http://conexp.sourceforge.net/users/documentation.html, 2014. [Online; accesat la 15-Iunie-2014].

[Zak11] Alon Zakai. Emscripten: An llvm-to-javascript compiler. In *Proceedings of the ACM International Conference Companion on Object Oriented Programming Systems Languages and Applications Companion*, SPLASH '11, pages 301–312, New York, NY, USA, 2011. ACM.

[Zak14] Alon Zakai. kripken/sql.js. https://github.com/kripken/sql.js/, 2014. [Online; accesat la 15-Iunie-2014].