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

Preface

As defined in [Wik14]

Big-data și machine learning sunt domenii foarte căutate în zilele noastre, devenind chiar “buzzwords”. Majoritatea se bazează pe algoritmi simplii, de adunare și prelucrare automată a datelor. Analiza conceptuală formală oferă o alternativă...

- Importanța explorării conceptelor
- Lacunele softului existent
- Intențiile aplicației

Lista acronimelor

HA = Harap Alb
PLL = Păsări-Lăți-Lungilă
Sp = Spânul

Capitolul 1
Analiza conceptuală formală

Analiza conceptuală formală este o metodă de sistematizare a datelor în **concepte**, definite la modul larg ca mulțimi de obiecte care împărtășesc anumite atribute sau proprietăți. Este o reintepretare 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.

1.1 Concepte matematice de bază

Întrucât scopul lucrării este de a descrie aplicația practică, ne vom limita la a descrie conceptele de care avem nevoie pentru a înțelege domeniul.

1.1.1 Mulțimi ordonate, latice, latice complete

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

Reflexivitate xRx

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

Tranzitivitate $xRy, yRz \Rightarrow xRz$

$\forall x, y, z \in M$. *Relația R se numește o **relație de ordine**.*

Cel mai simplu exemplu, intuitiv exemplu este mulțimea numerelor reale $\mathbb{R}$, alături de relația $\leq$. Notăm o mulțime ordonată cu relația $\leq$ cu $(M, \leq)$.

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

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

Definiție 3. *Două elemente ale unei mulțimi ordonate sunt **comparabile** 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. *Fie $(M, \leq)$ o mulțime ordonată, și N o submulțime a sa. Înțelegem prin **limita inferioară** a mulțimii N un element i astfel încât $\forall a \in N, i \leq a$. În mod invers, **limita superioară** a mulțimii s este definită prin $\forall a \in N, s \geq a$. Putem nota mulțimea tuturor limitelor inferioare a grupului N cu I. Elementul cel mai mare din această mulțime este numit **minorantul** mulțimii N. Invers, cel mai mic element din mulțimea limitelor superioare este numit **majorantul** mulțimii N.*

Minorantul se poate nota cu $\wedge N$ sau $\inf N$ (de la **infimum**), iar majorantul cu $\vee N$ sau $\sup N$ (de la supremum).

Definiție 5. *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 majorant și minorant. O latice este **completă** dacă pentru orice submulțime a ei există majorant și minorant.*

Orice latice completă are un element superior, numit **elementul unitate**, și un element inferior, numit **elementul zero**.

Definiție 6. *Conform [CR04] O **conexiune Galois** este compusă din două mulțimi ordonate, M, N și două funcții 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, $m \leq \psi n \Leftrightarrow n \leq \gamma m$

1.1.2 Context, concept, ierarhie de concepte

Definiție 7. *În cadrul analizei conceptuale, un **context** $K = (G, M, I)$ este format din 2 mulțimi, G și M, și o relație binară I între acestea. Mulțimea G reprezintă obiecte, iar M atribute.*

Literele provin din limba germană, în care conceptele au fost descrise inițial, de la Gegenstände și MerKmale, respectiv. Relația I e numită **relația de incidență**, iar gIm poate fi citit ca "obiectul g este descris de atributul m", sau "atributul m descrie obiectul g".

Exemplu 1. *Preluăm următorul exemplu din [CR04], un context (foarte redus)al animalelor vertebrate.*

		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 lu-mină (i)*
1	*Liliac*		×			×	×		×	
2	*Vultur*		×	×		×	×			
3	*Maimuță*				×	×			×	
4	*Pește papagal*	×		×		×		×		
5	*Pinguin*			×		×	×	×		
6	*Rechin*	×				×		×		
7	*Pește lanternă*	×				×		×		×

Tabela 1.1: Un conte×t al animalelor vertebrate.

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 8. *Un* ***concept*** *al contextului* (G, M, I) *este definit de* $A \subseteq G$, $B \subseteq M$, *unde* $A' = B$ *și* $B' = A$.

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

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. $A \subseteq B' \Longleftrightarrow B \subseteq A' \Longleftrightarrow A \times B \subseteq I$

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 extinderea unui concept. Din cele descrise mai sus, rezultă că e necesar ca $A = A''$ pentru A să fie extinderea 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 deasemenea și peștele lanternă, și peștele papagal.

- Intersecția oricâtor extinderi (obiective) de concepte rezultă întotdeauna o altă extindere (obiectiv).

- Reuniunea lor, pe de altă parte, rareori rezultă o altă extindere.

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

  Definiție 9. *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)

- Diagrame Hasse

- Algoritmi folosiți la generarea și afișarea laticelor.

1.2 Introducere

1.3 Teorie

1.4 Utilizări practice

Capitolul 2
Starea actuală

2.1 Toscana

2.2 Toscanaj

Capitolul 3
Romagna

3.1 Raționament

3.1.1 Ușurință de utilizare

3.1.2 Avantajele distribuirii aplicațiilor web

3.2 Structură

3.3 Tehnologii

3.3.1 CoffeeScript

3.3.2 Ember.js

3.3.3 d3.js

3.3.3.1 svg

3.3.4 sql.js

3.3.4.1 emscripten

3.4 Dezvoltare

3.4.1 Proces

3.4.2 Probleme întâmpinate

3.4.2.1 MySQL versus sql.js

Bibliografie

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

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

[Wik14] Wikipedia. Latex — wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=LaTeX&oldid=413720397`, 2014. [Online; accessed 14-June-2014].

[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.