Formal Concept Analysis applied to Professional Social Networks

Paula Silva, Sérgio M. Dias, Wladmir Brandão, Mark Song, Luis Zárate (supervisor)

Pontifical Catholic University of Minas Gerais, Brazil

ICEIS, 2017

Context and Motivation

- Online social networks for users oriented to business
- The LinkedIn is the one of the largest and most popular online professional social network
- The size and diversity of users generated content data
- The need to help professionals to increase skills and reach job positions
- The Formal Concept Analysis (FCA) as mathematical formulation for data analysis, applied to find patterns of professional competence

Goal

- Identify professional behaviors through data scraped from LinkedIn
- Find the minimum set of skills that is necessary to reach job positions
    - For example: statistic, machine learning, databases $\rightarrow$ data scientist
- Implications rules, specifically the set of proper implications

Contributions

- Our contributions are:
  - The domain problem mapping the model of competences
  - The professionals data set scraped from LinkedIn
  - The FCA-based approach
  - The set of experiments to apply FCA for professional career analysis

Formal Concept Analysis

- Based on the notions of concept and conceptual hierarchy
- A mathematical way to look at data and knowledge, their acquisition process and analysis based on lattices.
- There are three main principles:
  - Formal context
  - Formal concept
  - Implications

Formal Context

- Formal context (G, M, I)
  - a set G of objects
  - a set M of attributes
  - a binary relation $I \subseteq G \times M$

	a	b	c	d	e	f	g
18			x	x			
19		x				x	x
20	x	x		x	x		
21			x	x			x
22		x				x	
23	x	x		x	x		x
24			x	x			

Table: Example context of an user's *LinkedIn* skills.

Formal Concept

- Derivation operators:
  For $A \subseteq G$ and $B \subseteq M$

$$A' := \{m \in M | gIm \forall g \in A\}$$

$$B' := \{g \in G | gIm \forall m \in B\}$$

- Formal concept (A, B)

$$\begin{array}{cc} A \subseteq G & B \subseteq M \\ A' = B & B' = A \end{array}$$

  A is concept extent and B is concept intent

- e.g. formal concept $(\{18, 21, 24\}, \{$software engineering, data bases$\})$

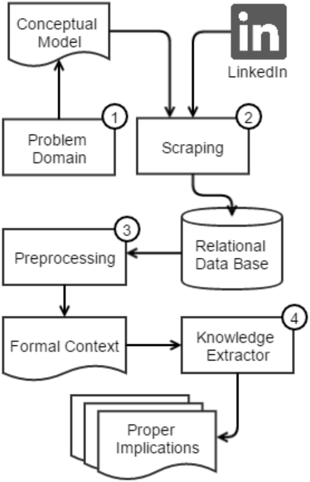
Implication Rules

- Being a formal context (G, M, I), an implication over M is $P \rightarrow Q$, where $P, Q \subseteq M$.
- P is the premise
- Q is the conclusion
- $P \rightarrow Q$ has to be such that $P' \subseteq Q'$, so the sets of attributes P and Q share the same subset of objects.

The Set of Proper Implications

- the right hand side of each implication is unitary: if $P \to m \in \mathscr{I}$, then $m \in M$;
- superfluous implications are not allowed: if $P \to m \in \mathscr{I}$, then $m \notin P$;
- specializations are not allowed, i.e. left hand sides are minimal: if $P \to m \in \mathscr{I}$, then there is not any $Q \to m \in \mathscr{I}$ such that $Q \subset P$.
- For example, $\{e\} \to \{a\}$ is a proper implication, but $\{e, g\} \to \{a\}$ is not a proper implication.

FCA-based Approach



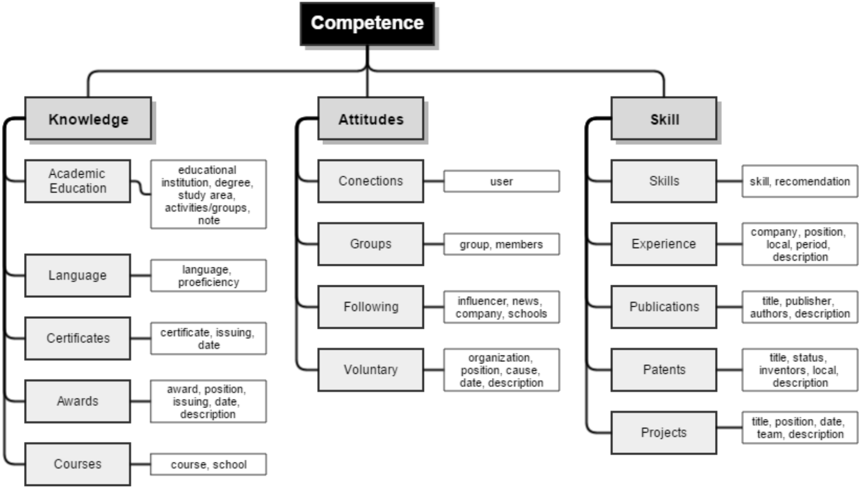
1 - Problem Domain

- Building the conceptual model according to the problem to be treated
- The classification of informational categories, based on Model of Competence

Figure: Duran(1998) adapted model

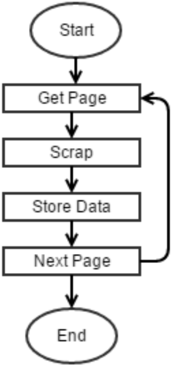
1 - Problem Domain

- We identified 3 dimensions, 14 aspects and 51 variables



2 - Scrapping

- The scrapping process was divided into two phases.
  - Selecting initial seeds randomly
  - Collecting the public profiles data: People undergraduate in IT coursers in Minas Gerais, Brazil



3 - Preprocessing

- We only considered the variables *skills* and *experience*
- The ETL process:
  - String cleaning: UTF-8 encoding correction, accent removal, standardization of all terms for the English language through Google Translate API
  - Attribute reductions: reductions based on semantic relevance

Generic term	Specific term
Java	JPA JSF
Software developer	Developer Programmer Program developer

  - Formal context with 366 attributes and 970 objects

4 - Knowledge Extractor

- The PropIm algorithm
  - Finding supersets: If $A \rightarrow b$, so $\uparrow |(A \rightarrow b)'|$ and $\downarrow |A|$
  - It computes proper implications with support > 0
  - Easily scalable strategy
  - Allows to set the attributes of interest for implications' conclusions
  - Problem complexity: $O(|M||\mathscr{I}|(|G||M| + |\mathscr{I}||M|))$
  - Pruning heuristic to reduce combinations possibilities among attributes

4 - Knowledge Extractor

Input : Formal context (G, M, I)
Output: Set of proper implications $\mathscr{I}$ with support greater than 0

$\mathscr{I} = \emptyset$
foreach $m \in M$ **do**
  $P = m''$
  $size = 1$
  $Pa = \emptyset$
  while $size < |P|$ **do**
    $C = \binom{P}{size}$
    $P_C = getCandidate(C, Pa)$
    foreach $P1 \subset P_C$ **do**
      if $P1' \neq \emptyset$ *and* $P1' \subset m'$ **then**
        $Pa = Pa \cup \{P1\}$
        $\mathscr{I} = \mathscr{I} \cup \{P1 \rightarrow m\}$
      end
    end
    $size{+}{+}$
  end
end
return $\mathscr{I}$

Function *getCandidate* (C, Pa)
  $D = \emptyset$
  foreach $a \in A | A \subset Pa$ **do**
    foreach $B \subset C$ **do**
      if $a \notin B$ **then**
        $D = P_C \setminus B$
      end
    end
  end
  return D

Experiments

The goal was to answer the following questions:

- How do proper implications identify relations between skills and positions?
- Could we find intersections among sets of skills, and what do these intersections represent?

Proper Implications to Competence Identification

- We selected 20 positions and their 180 skills
- It was extracted 895 proper implications with PropIm algorithm

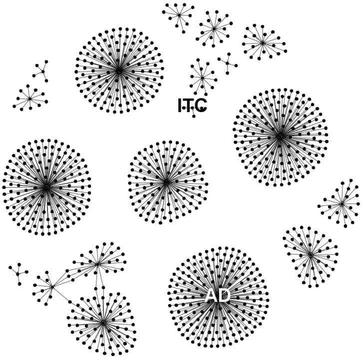


Figure: Proper implications network. AD = *Administrative Director*, ITC = *IT Consultant*

Proper Implications to Competence Identification

- Nodes with high in-degree value represent positions which have more diversification of sets of skills
- 163 proper implications
- $\{$*entrepreneurship, human resources, information management*$\} \rightarrow \{$*administrative director*$\}$

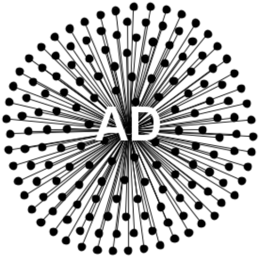


Figure: AD = *Administrative Director*

Proper Implications to Competence Identification

- Nodes with low in-degree value represent jobs positions that demand more specific sets of skills
- 3 proper implications
- $\{ABAP, agile\ methodology, BI\} \rightarrow \{it\ consultant\}$

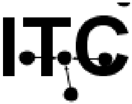


Figure: ITC = *IT Consultant*

Intersection Between Skills and Job Positions

- Top 3 best jobs in Information Technology area according to *Career Cast research* (Cast, 2016)
- Edges weight = relative frequency $\mathscr{F} = \dfrac{F_i}{F_p}$, where $\mathscr{F}$ is the relative frequency, F_i is the implication absolute frequency and F_p is $|m'|$
- Why relative frequency?
  $\{$java frameworks$\} \rightarrow \{$software engineer$\}$
  Support: 2.47%
  Relative frequency: 75.56%

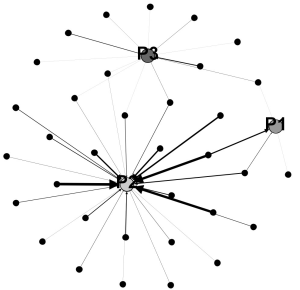


Figure: P_1: *data scientist*, P_2: *information security analyst* and P_3: *software engineer* job position

Intersection Between Skills and Job Positions

- Nodes P_1 and P_2 share two set of skills:
  - $\{$*agile methodology*$\} \rightarrow \{$*data scientist*$\}$
  - $\{$*agile methodology*$\} \rightarrow \{$*information security analyst*$\}$

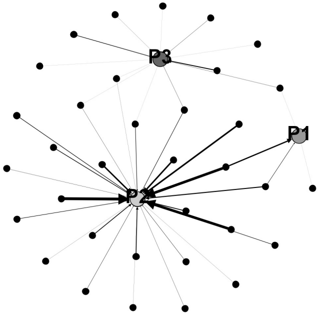


Figure: P_1: *data scientist*, P_2: *information security analyst* and P_3: *software engineer* job position

Intersection Between Skills and Job Positions

- Skills and positions related to hierarchical transition:
  - P_1: $\{$*.NET, automation systems*$\} \rightarrow$ *IT analyst*
  - P_2: $\{$*.NET, data base, ERP, it governance*$\} \rightarrow$ *IT coordinator*
  - P_3: $\{$*BPM, cloud computing, CRM*$\} \rightarrow$ *IT manager*
  - P_4: $\{$*assets management, BI, business management, consulting*$\} \rightarrow$ *IT director*

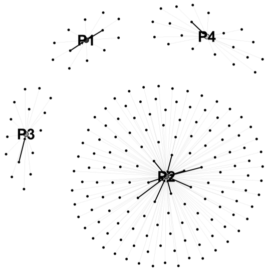


Figure: P_1 to P_4 represents the following job positions: P_1 is *IT analyst*, P_2 is *IT manager*, P_3 is *IT coordinator* and P_4 is *IT director*

Conclusions

- FCA-based approach to identify the minimum sets of skills that is necessary to achieve a job position
- Set of experiments for apply FCA to professional competences analysis
- Computational strategy to find proper implications without loss of information
- In-degree of conclusions nodes mean the diversification among sets of skills for the same job position
- Sharing sets of skills do not determine the job position, but show positions that can be achieved
- The disjointed sets (representing hierarchy) show that is necessary develop skills of different natures to progress in career

Future Works

- Experiments will be replicated for other areas
- Exploring another algorithms which extract implications from concept lattice, or from the set of formal concepts
- Expanding the analysis to all dimensions from *model of competence*
- Implementing an web environment with this FCA-based approach, for help professionals to increase skills and look for possible job positions
- Preprocessing: attribute fusion through correlation analysis
- Temporal analysis of career evolution
- Estimate the remuneration of people based on posts and professional data

Any questions?

paula.raissa@sga.pucminas.br
paula.csraissa@gmail.com