Formal Concept Analysis Applied to Professional Social Networks Analysis

Paula R. C. Silva[1], Sérgio M. Dias[1,2], Wladmir C. Brandão[1], Mark A. Song[1] and Luis E. Zárate[1]
[1]*Informatics Institute, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil*
[2]*Federal Service of Data Processing, Belo Horizonte, Brazil*

Keywords: Formal Concept Analysis, Proper Implications, Social Networks.

Abstract: From the recent proliferation of online social networks, a set of specific type of social network is attracting more and more interest from people all around the world. It is professional social networks, where the users' interest is oriented to business. The behavior analysis of this type of user can generate knowledge about competences that people have been developed in their professional career. In this scenario, and considering the available amount of information in professional social networks, it has been fundamental the adoption of effective computational methods to analyze these networks. The *formal concept analysis (FCA)* has been a effective technique to *social network analysis (SNA)*, because it allows identify conceptual structures in data sets, through conceptual lattice and implication rules. Particularly, a specific set of implications rules, know as proper implications, can represent the minimum set of conditions to reach a specific goal. In this work, we proposed a FCA-based approach to identify relations among professional competences through proper implications. The experimental results, with professional profiles from *LinkedIn* and proper implications extracted from *PropIm* algorithm, shows the minimum sets of skills that is necessary to reach job positions.

1 INTRODUCTION

In increasingly interconnected world, the social networks attend different people's interests and address the communication and information needs of several user groups (Russell, 2013). In particular, there are professional social networks focused on a specific group of users interest is oriented to business. One of the largest and most popular professional social network is *LinkedIn*, which has more than 400 millions users distributed in more than 200 countries and territories (LinkedIn, 2016).

LinkedIn users create professional profiles, they available their professional skills, competences and experience. Thus, *LinkedIn* provides a source of professional information that can be exploited by enterprise managers in different ways, e.g., to find people with appropriate competences to fulfill specific positions. In addition, the size and diversity of user-generated content create an opportunity to identify behavioral trends and user communities. In this scenario, the *formal concept analysis (FCA)* presents itself as a mathematical theory that can be used for this purpose.

FCA presents a mathematical formulation for data analysis, which identify conceptual structures from a data set (Ganter et al., 2005). It also presents an interesting unified framework to identify dependencies among data, by understanding and computing them in a formal way (Codocedo et al., 2016). It is a branch of lattice theory motivated by the need for a clear formalization of the notions of concept hierarchy.

There are two ways to extract and represent knowledge from FCA: conceptual lattice and implication rules. In this work, we applied a particularly type of implication rules, know as *proper implications* (Taouil and Bastide, 2001). We say that a proposition P logically implies a proposition Q ($P \rightarrow Q$), if Q is true whenever P is true. The set of proper implications have a minimal left-hand side and only one item in right-hand side (Taouil and Bastide, 2001). It has been used when the need is to find the minimum conditions to lead a goal. In this article, the proper implications represent the set of minimum professional's skills (conditions) for achieve a job position (goal). For example, the proper implication $\{$*statistic, machine learning, databases*$\} \rightarrow \{$*data scientist*$\}$ represent a minimum set of skills which are necessary to be a data scientist.

Several authors have been applied FCA to address

Silva, P., Dias, S., Brandão, W., Song, M. and Zárate, L.
Formal Concept Analysis Applied to Professional Social Networks Analysis.
DOI: 10.5220/0006333401230134
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 123-134
ISBN: 978-989-758-247-9

research problems related to *social networks analysis (SNA)*. Note that, there are other methods to retrieve knowledge from social networks. They are usually based on graph theory, clustering and frequent itemsets, which provide an approach to represent a social network through a formal way. Additionally, there are several FCA to SNA applications, such as ontology-based technique (Kontopoulos et al., 2013), social communities (Ali et al., 2014), network representation through concept lattice (Cuvelier and Aufaure, 2011), contextual pre-filtering process and identifying user behavior through implication rules (Neto et al., 2015a).

In this article, we proposed a FCA-based approach to identify professional behaviors through data scraped from *LinkedIn*. First, we conceptually model a *LinkedIn* user profile according to the *model of competences* proposed by Durand (1998), because this model has more acceptance in industry and academia (Brandão and Guimarães, 2001). Second, as an input data set, our approach needs an incidence table (formal context) and a subset of proper implications are expected as an output. Lastly, a proper implication represent careers trajectories, by minimum professional's skills (conditions) for achieve a position (goal). These implications was extracted using our proposed algorithm named *PropIm*. The *PropIm* algorithm was proposed to extract proper implications with support greater than zero, using a scalable approach.

The main contributions of this paper are: a professionals data set scraped from *LinkedIn*, a FCA-based approach, and experiments set for apply FCA to professional career analysis.

The structure of the article follows as: In Section 2, we present the preliminary definitions related to FCA. In Section 3, we report related work that applied FCA to social networks analysis. In Section 4, we present our FCA-based approach to SNA. In Section 5, we report experiments and results. Finally, in Section 6, we present the conclusions and future work.

2 FORMAL CONCEPT ANALYSIS

In this section, we introduce concepts related of formal concept analysis (FCA) reported on literature (Ganter and Wille, 2012).

2.1 Formal Context

Formally, a formal context is a triple (G,M,I), where G is a set of objects (rows), M is a set of attributes (columns) and I are incidences. It is defined as $I \subseteq G \times M$. If an object $g \in G$ and an attribute $m \in M$ have a relationship I, their representation is $(g,m) \in I$ or gIm, which could be read as "object g has attribute m". When an object has an attribute, the incidence is identified and represented by "x". In the formal context shown in Table 1, rows are objects representing users, and the columns are professional skills and positions.

Table 1: Example context of an user's *LinkedIn* skills. The attributes are: a: networks, b: mobile application, c: software engineering, d: data bases, e: graphic processing, f: computer architecture, g: operational systems.

	a	b	c	d	e	f	g
17	x	x		x	x		x
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			

Given a subset of objects $A \subseteq G$ of formal context (G,M,I), there is an attribute subset of M common to all objects of A, even if empty. Likewise, given a set $B \subseteq M$, there is an object subset that shares the attributes of B, even if empty. These relationships are defined by derivation operations:

$$A' := \{m \in M | gIm \forall g \in A\} \quad (1)$$

$$B' := \{g \in G | gIm \forall m \in B\} \quad (2)$$

A formal context (G, M, I) is a clarified context, when $\forall$ g, $h \in G$, from $g' = h'$ it always follows that and, correspondingly $m' = n'$ implies $m = n$ $\forall$ $m,n \in M$. The clarification process consists in maintaining one element (objects and attributes) from a set of equal elements eliminating the others. In this process, the number of objects and attributes can be reduced while retaining lattice form (Ganter et al., 2005).

2.2 Formal Concept

From formal contexts we can obtain formal concepts, defined as pairs (A,B), where $A \subseteq G$ is called extension and $B \subseteq M$ and is called intention, and they must follow conditions $A = B'$ and $B = A'$ (equations 1 and 2)(Ganter et al., 2005).

Based on the formal context from Table 1, we can obtain the formal concept ({18, 21, 24}, {software engineering, data bases}), where elements of subset B are {software engineering, data bases}, that, by derivation (Equation 2), yield subset A = {18, 21,

24}, representing the subset of users who have skills in *software engineering (c)* and *data bases (d)*.

It is important to note that a formal concept corresponds to any aspect of the problem domain, represented by objects and attributes, in which exists some kind of comprehension and understanding.

2.3 Concept Lattice

With all formal concepts sorted hierarchically by order of inclusion $\subseteq$, we can build the concept lattice. Sorting must be done, so that, the concept (A_1, B_1) is considered less than or equal to (A_2, B_2) if and only if, $A_1 \subseteq A_2$ (equivalent to $B_2 \subseteq B_1$). In this case, the concept (A_1, B_1) is called sub-concept and the concept (A_2, B_2) super-concept. In Figure 1 is shown an example of a concept lattice from the formal context in Table 1.

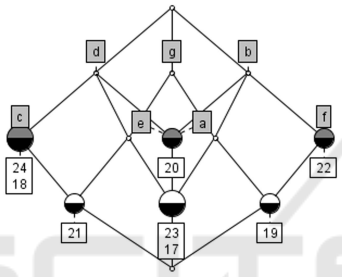

Figure 1: Example of concept lattice.

2.4 Implication Rules

Given a formal context (G, M, I) or a concept lattice $\mathscr{B}(G, M, I)$, these can be extracted exact rules or approximate rules (rules with statistical values, for example, support and confidence) that express in a alternative way the underlying knowledge. The exact rules can be classified in implication rules and functional dependencies, while the approximation rules are divided in classification rules and association rules. It is particularly important in this work, to get the social networks users' behavior, consider exact rules. From now these rules are going to called only *implications*. Follows the definition of an implication (Ganter and Wille, 2012):

Definition 2.1. *Being a formal context whose attributes set is M. An implication is an expression* $P \to Q$*, which* $P, Q \subseteq M$.

An implication $P \to Q$, extracted from a formal context, or respective concept lattice, have to be such that $P' \subseteq Q'$. In other words: every object wich has the attributes of P, it also have the attributes of Q.

Note that, if X is a set of attributes, then X *respects* an implication $P \to Q$ iff $P \not\subseteq X$ or $Q \subseteq X$. An implication $P \to Q$ *holds* in a set $\{X_1, \ldots, X_n\} \subseteq M$ iff each X_i respects $P \to Q$; and $P \to Q$ *is an implication of the context* (G, M, I) iff it holds in its set of object intents (an object intent is the set of its attributes). An implication $P \to Q$ *follows from* a set of implications $\mathscr{I}$, iff for every set of attributes X if X respects $\mathscr{I}$, then it respects $P \to Q$. A set of implications $\mathscr{I}$ is said to be *complete* in (G, M, I) iff every implication of (G, M, I) follows from $\mathscr{I}$. A set of implications $\mathscr{I}$ is said to be *redundant* iff it contains an implication $P \to Q$ that follows from $\mathscr{I} \setminus \{P \to Q\}$. Finally, an implication $P \to Q$ is considered *superfluous* iff $P \cap Q \neq \emptyset$.

In social networks will be convenient that each implication represent a minimum behavior. For this, we will require that the complete set of implications $\mathscr{I}$ of a formal context (G, M, I) have the following characteristics, to be used as representative of the process:

- 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 $Q \subset P$.

A complete set of implications in (G, M, I) with such properties is the so called set of *proper implications* (Taouil and Bastide, 2001) or *unary implication system* (UIS) (Bertet and Monjardet, 2010).

Definition 2.2. *Let* $\mathscr{J}$ *be the complete closed set of implications of a formal context* (G, M, I)*. Then the set of* proper implications $\mathscr{I}$ *for* (G, M, I) *is defined as:* $\{P \to m \in \mathscr{J} \mid P \subseteq M$ *and* $m \in M \setminus P$ *and* $\forall Z \subset P : Z \to m \notin \mathscr{J}\}$.

Table 2: Proper implications extracted from formal context in Table 1.

$P \to m$	
e b,d	$\to$ a
a e f	$\to$ b
d,g	$\to$ c
a c e	$\to$ d
a b,d	$\to$ e

The Table 2 shows the set of proper implications, from the formal context example (Table 1). For example, in implication $b,d \to a$, the set P is composed by the set of attributes $\{b, d\}$, $m = \{a\}$ and $\to$ symbol represents the incidence. P and m are called as

premise and conclusion. So the implication $b,d \rightarrow a$, can be read as the premise b,d implies in conclusion a. It is important to note that, the conclusion a can has more than one minimal premises. According to Definition 2.2, the premises $\{e\}$ and $\{b, d\}$ are minimal, and they can not be a subset of another premises that imply in a. For example, the implication $e \rightarrow a$ is a proper implication, but $e, g \rightarrow a$ is not a proper implication.

3 RELATED WORK

Recently, several authors have applied FCA for social networks analysis (Rome and Haralick, 2005; Snasel et al., 2009; Stattner and Collard, 2012; Aufaure and Le Grand, 2013; Banerjee et al., 2014; Krajči, 2014; Atzmueller, 2015; Neznanov and Parinov, 2015; Soldano et al., 2015). These works have been motivated by the interest in understand and interpret social networks through mathematical formulation. The main subjects are: social network representation as a concept lattice, community detection, concepts mining, ontology analysis and rule mining through implications.

In (Cordero et al., 2015), the authors proposed knowledge-based model of influence applying FCA to compute minimal generators and closed sets directly from an implicational system, for obtain a structure of user's influence. The data was extracted from *Twitter* social network, it was transformed into a formal context and it was generated the Duquenne-Guigues basis. In (Neto et al., 2015b), the author shows an approach to analyze a data base composed by internet's access logs. The authors apply the minimal set of implications and complex networks theories to identify substructures, that are not easily visualized with two-mode networks. In (Jota Resende et al., 2015), the authors propose an FCA-based approach to build canonical models, which represents *Orkut* access' patterns. These papers resemble this work, because they also talks about how to map social networks in terms of objects and attributes, and extract knowledge through implications rules set.

As this work, in Barysheva et al. (2015) the authors apply FCA to identify interaction patterns through data scraped from *LinkedIn*. They did not work with professional competencies and proper implications, like us, because their goal was classify users' behavior though their network interactions, and the knowledge was extracted from conceptual lattice.

The papers of Li et al. (2016); Xu et al. (2014a); Lorenzo et al. (2016) are the main works about career trajectory analysis using *LinkedIn* data. Their goal was model professional profiles and identify career trajectories through temporal analysis. Even not applying FCA, these authors apply data mining techniques to identify professional career trajectories.

In general, the FCA has been applied to mining social media, because the FCA theory presents a formalism for the representation of network structure, behavior identification and knowledge extraction, through formal representation of problem domain from objects, attributes and their respective incidences.

Our proposed approach combines techniques conducted on formal concept analysis, patter mining and model of professional competence.

4 FCA-BASED APPROACH

In this article, the problem of analysis and representation of professional profiles in social networks, can be grounded by building a conceptual model, that merge the social network with professional skills theory, and the transformation of this model to a formal context. After scraping and pre-process the data to a formal context, we can extract the set of implications to be analyzed. The Figure 2 shows the methodology steps proposed to analyze *LinkedIn* social network through FCA.

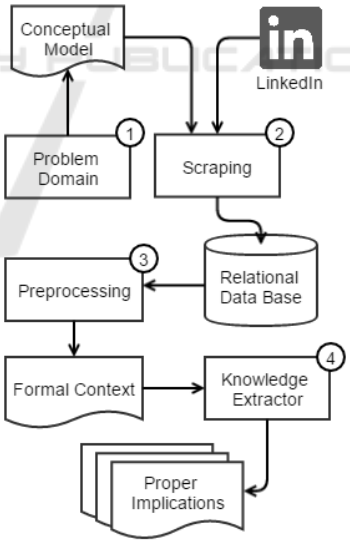

Figure 2: Methodology based in FCA to SNA.

4.1 Problem Domain

According to Figure 2, the first step (1) was to construct the conceptual model according to the problem to be treated. In this case, the problem involves the characterization of a person as a professional. We

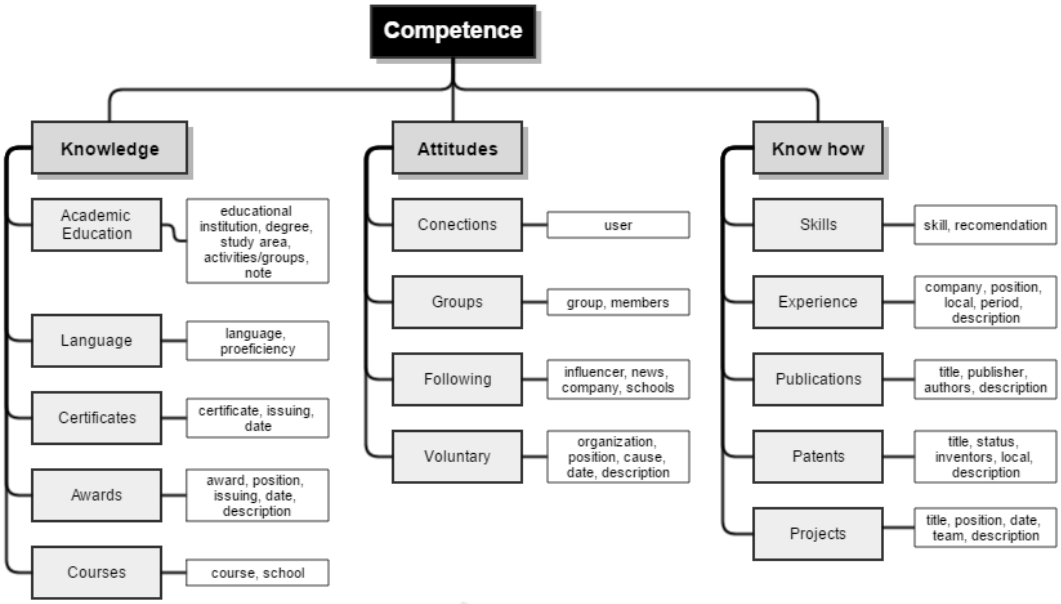

Figure 3: Professional identification through model of competence.

adopted the *competence model* proposed by Durand (1998), because this model has greater acceptance in both academic and business, since it seeks to integrate aspects related to this work, such as technical issues, cognition and attitudes (Brandão and Guimarães, 2001). For Durand (1998), a professional is characterized by his competence in accomplishing a certain purpose. Competence is composed by three dimensions: *knowledge*, *attitudes* and *know how*. The dimensions are interdependent, because the behavior of a professional is determined not only by his *knowledge* but also by the *attitudes* and *know how* acquired over time. The *knowledge* dimension is linked to academic training and complementary courses. The *know how* dimension is related to a person's professional experiences. And finally, the *attitude* dimension is related to the way of people interact in their professional environment.

The conceptual model was created from the classification of informational categories, based on the model of competence. Figure 3 shows the domain model for professional identification through model of competence. The first level is related to the *Competence* concept, in the second level is 3 dimensions, in the third level is 14 aspects and in the fourth level is 51 variables.

4.2 Scrapping

FCA techniques to social network analysis can yield insights into user behaviors, detecting popular topics, and discovering groups and communities with similar characteristics. So, a task of gathering the data on a specific subject is needed. In this case, the second step (2) of the methodology represents the Scraping component that is responsible for collecting the *LinkedIn* user data.

The collection process was divided into two phases. The first one selects the initial seeds, randomly two user profiles were selected. This amount of initial seed was satisfactory for the data collection process, due to the total of profiles obtained being sufficient for the study. It was defined that, as case study, the data would be collected from people of Belo Horizonte, Minas Gerais, Brazil and they must have at least graduate courses in the information technology (IT) area.

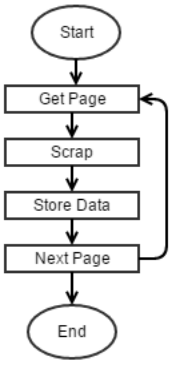

Figure 4: Scraping *LinkedIn* profile's process.

The second phase goal is collect the public profiles. As the *LinkedIn* does not provide an API (Application Programming Interface) to extract data directly

from the server, an approach, known as open collection, has been adopted to extract data from users' public pages.

The Figure 4 exemplifies the flow of data collection process. The process starts by accessing a seed. In the public profile there is a section denoted as "People also saw" - a list with the 10 most similar profiles related to the visited profile (Xu et al., 2014b). The collector looks up these addresses and verifies which profiles meet the scope. Valid profiles are stored and each one becomes a new seed to extract new links, restarting the collection process until it reaches the stop criterion. The stopping criterion is based on the percentage of new profiles. Each iteration checks if 65% of the profiles were already in the database.

4.3 Preprocessing

The Preprocessing (3) component is responsible for pre-processing the data extracted in the previous step. In this work only the variables *skills* and *experience* were considered. However, in future works, the other dimensions will be included.

For the construction of the formal context, we considered, as attribute, each value of the *competence* and *experience* variables. Each user (professional of *LinkedIn*) is considered as an object. In the first version of the formal context, 4000 attributes and 1280 objects were detected. As such values are texts and *LinkedIn* allows users to fill the corresponding fields with a free text, some problems have been detected. In this case, it was necessary to create an ETL process (Extract Transform Load) to clean and transform the data, aiming to reduce the amount of attributes.

The ETL process consists of two stages. In the first step, we applied basic procedures for string cleaning, like: UTF-8 encoding correction, accent removal, and standardization of all terms for the English language through Google Translate API [1]. In the second stage, we apply techniques to attribute reductions. Such reductions were based on terms with semantic relevance, in which the attributes *skill* and *experience* could be reduced. For example, attributes as *jpa*, *jsf* were renamed to *java frameworks*; attributes as *developer*, *programmer*, *software developer*, *program developer* were renamed to *software developer*. The vague nouns or trademark terms, as *bachelor*, *engineer*, *accessibility*, *microsoft*, were removed, because they are not relevant to our study.

At the end of the preprocessing step, a formal context was created with 366 attributes and 970 objects, which 61 attributes are related to *experience* and 305 to *skills*.

[1]Translate API: https://cloud.google.com/translate/

4.4 Knowledge Extractor

The *Impec* (Taouil and Bastide, 2001) is the best-known algorithm for extracting proper implications from the formal context. This is due to the strategy that the algorithm adopts at the moment in which it finds the premises and their respective conclusions.

On the other hand, the strategy used, by the algorithm, is computationally inefficient and it can not be optimized or distributed. In some cases, only a few context attributes are interesting to be in the conclusion. The traditional algorithms only allow to generate the complete set, being necessary a step of filtering the rules to select to those whose attributes of interest appear as conclusion of the implications, occurring an unnecessary computational effort.

Based on the problems described above, a new algorithm was proposed to generate proper implications from the formal context. The algorithm adopts an easily scalable strategy and allows its proper implications to be generated from the attributes of interest as a conclusion of implications. In general, the algorithm receives a formal context as input, finds the minimum premisses for each conclusion by combining attributes, applies a pruning heuristic, and uses the derivation operators to validate the implications.

Input : Formal context (G,M,I)
Output: Set of proper implications $\mathscr{I}$ with support greater than 1

1 $\mathscr{I} = \emptyset$
2 **foreach** $m \in M$ **do**
3 $P = m''$
4 $size = 1$
5 $Pa = \emptyset$
6 **while** $size < |P|$ **do**
7 $C = \binom{P}{size}$
8 $P_C = getCandidate(C, Pa)$
9 **foreach** $P1 \subset P_C$ **do**
10 **if** $P1' \neq \emptyset$ *and* $P1' \subset m'$ **then**
11 $Pa = Pa \cup \{P1\}$
12 $\mathscr{I} = \mathscr{I} \cup \{P1 \rightarrow m\}$
13 **end**
14 **end**
15 $size++$
16 **end**
17 **end**
18 **return** $\mathscr{I}$

Algorithm 1: Extract proper implications with support greater than 1.

The pseudo-code of *PropIm* is given in Algorithm 1. The main objective is to find implications whose left side is minimal and the right side has only one attribute. The algorithm needs a formal context

Table 3: Example of PropIm algorithm execution.

m	*P*	*size*	*C*	*Pc*	*Pa*	$\mathscr{I}$
a	{b, d, e, g}	1	{{b}, {d}, {e}, {g}}	{{b},{d}, {e}, {g}}	{{e}}	{{e-a}}
a	{b,d,e,g}	2	{{bd}, {be}, {bg}, {de}, {dg}, {eg}}	{{bd},{bg}, {dg}}	{{e}, {bd}}	{{e-a}, {bd-a}}
a	{b,d,e,g}	3	{{bde}, {bdg}, {beg}, {deg}}	∅	{{e},{bd}}	{{e-a},{bd-a}}
a	{b,d,e,g}	4	{{b,d,e,g}}	∅	{{e}, {bd}}	{{e-a},{bd-a}}
b	{a,d,e,f,g}	1	{{a}, {d}, {e}, {f}, {g}}	{{a},{d}, {e},{f},{g}}	{{a}, {e}, {f}}	{{e-a}, {bd-a}, {a-b}, {e-b}, {f-b}}
b	{a,d,e,f,g}	2	{{ad}, {ae}, {af}, {ag}, {de}, {df}, {dg}, {ef}, {eg}, {fg}}	{{dg}}	{{a}, {e}, {f}}	{{e-a}, {bd-a}, {a-b}, {e-b}, {f-b}}
b	{a,d,e,f,g}	3	{{ade}, {adf}, {adg}, {aef}, {aeg}, {afg}, {def}, {deg}, {dfg}, {efg}}	∅	{{a}, {e}, {f}}	{{e-a}, {bd-a}, {a-b}, {e-b}, {f-b}}
c	{d,g}	1	{{d}, {g}}	{{d},{g}}	∅	{{e-a}, {bd-a},{a-b}, {e-b}, {f-b}}
c	{d,g}	2	{{dg}}	{{dg}}	∅	{{e-a}, {bd-a}, {a-b}, {e-b}, {f-b}}

(G,M,I) as input, and its output is a set of proper implications. Line 1 initializes the set $\mathscr{I}$ with empty set. The following loop (Lines 2-17) looks at each attribute in the set M. We initially suppose that each attribute m can be a conclusion for a set of premises. For each m, we compute a left-hand side $P1$.

To reduce the searching space, the algorithm finds the right side P for a left side m from a set of attributes common to m objects. After, it finds sets of possible premises for m based on P. The *size* counter determines the size of each premise, as the smallest possible size is 1 (an implication of type $\{b\} \rightarrow \{a\}$), it is initialized with 1 (Line 4).

A set of auxiliary premises *Pa* is used, where all valid premises found leading to m conclusion are stored (at Line 5 *Pa* is initialized as empty). In the loop, from Lines 6-16, the set of minimum premises is found and is bounded by $|P|$. In Line 7, the C set gets all combinations of size *size* from elements in P. In Line 8, the set of candidate premises is formed through the function *getCandidate* which will be described next.

Each candidate premise $P1 \subset P_C$ is checked to ensure if the premise $P1$ and the conclusion m results in a valid proper implication. Case $P1 \neq \emptyset$ and $P \subset P1'$, the premise $p1$ is added to the set of auxiliary premises *Pa* and $\mathscr{I} = \mathscr{I} \cup \{P1 \rightarrow M\}$.

A neighborhood search heuristic was implemented by the *getCandidate* (Pseudo-code in Algorithm 2) function. The objective is to find, in the set of C combinations, all subsets B that do not contain some attribute that already belongs to some valid premise of *Pa*. It receives, as parameter, the sets C and *Pa*, and returns a set D of proper premises.

```
1  Function getCandidate (C, Pa)
2      D = ∅
3      foreach a ∈ A|A ⊂ Pa do
4          foreach B ⊂ C do
5              if a ∉ B then
6                  D = P_C \ B
7              end
8          end
9      end
10     return D
```

Algorithm 2: Function getCandidate.

Table 3 shows the steps of *PropIm* algorithm, on the example from Table 1. $\mathscr{I}$ contains initially $\emptyset$. The first value to m is a (first attribute from formal context) and m'' is the set of attributes $\{b,d,e,g\}$. The *size* of combination sets is 1, so $C = \{\{b\},\{d\},\{e\},\{g\}\}$. *Pa* contains initially $\emptyset$, $C \setminus \emptyset$, so the set of attributes returned by the function *getCandidate* is $Pc = \{\{b\},\{d\},\{e\},\{g\}\}$. For each subset of *Pc*, only the element $\{e\}$ attends the condition in Line 10 (Algorithm 1), because $\{e\}' = \{17,20,23\} \subset m'$. The set $\{e\}$ is added to *Pa* and the pair $\{e \rightarrow a\}$ is added to $\mathscr{I}$. When *Pc* is $\emptyset$ and *size* is $|P|$ the loop to $m = a$ is closed. So, the same steps happens for all attributes, from formal context, imputed to m.

5 EXPERIMENTS

This section shows the procedures, adopted for running the experiments, and the analyses of results obtained based in proposed FCA-based approach. The

goal, of experiments, 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?

5.1 Proper Implications to Competence Identification

Among the 61 positions identified in Section 4.3, we selected 20 positions and their 180 skills for analyze the proper implications. In this case, the *PropIm* algorithm extracted 895 proper implications related to this reduced formal context. Figure 5, shows these proper implications as a graph representation (proper implications network). The central nodes are the positions and the edges represents the implications between premises (set of skills) and their conclusion (position). In this study case, the graph representation helps us to analyze the distribution among sets of skills and their respective positions.

In Figure 5 was highlighted some items related to graph analysis:

- Central nodes density;
- Intersections among premises;
- Edges weight.

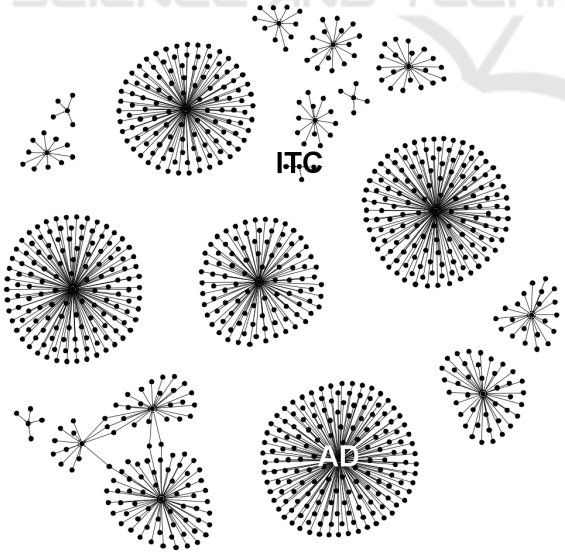

Figure 5: Proper implications network.

The central nodes density represents the diversification of minimum sets of skills. The denser nodes represents positions which have more diversification of minimum sets of skills. For example, the highlighted node *AD* related to *administrative director* position have 163 minimal sets of skills. Generally, the *administrative director* function is manage the company resources, like human, technologies and financial resources. The specific skills of this professional can be different according to the company industry, because he have to develop business skills and know how about the company resources. It is expected that *administrative director* develop skills related to leadership, management, technology and communication. One of the proper implications which represents this professional profile is $\{$*entrepreneurship, human resources, information management*$\} \rightarrow \{$*administrative director*$\}$. Another implication as $\{$*assets management, it governance, leadership development, software development*$\} \rightarrow \{$*administrative director*$\}$, can represent a specific *administrative director* from companies focused on software development.

The less dense nodes represent jobs positions that demand more specific sets of skills. For example the *ITC (IT consultant)* node have only 3 sets of skills related to it. An *IT consultant* duties can vary depending on the nature of company's project and client desires. However, in general, this professional has skills which combine IT and business knowledge. So, the proper implication $\{$*ABAP*[2]*, agile methodology, BI*[3]$\} \rightarrow \{$*it consultant*$\}$ shows the common set of skills that the IT consultants have.

The analysis related to intersection between premises and edges weight will be discussed in the next section (5.2).

5.2 Intersection between Skills and Job Positions

According to *Career Cast* research (Cast, 2016), the top 3 best jobs in *Information Technology* area is: data scientist, information security analyst and software engineer. From the set of proper implications, generated by *PropIm* algorithm, we filtered the *top 3* jobs positions, for analyze these jobs and identify the intersection among their skills.

Figure 6 shows the top 3 job positions and the intersections among their skills. The central nodes are the top 3 positions, according to *Career Cast* ranking (Cast, 2016): P_1 (*data scientist*), P_2 (*information security analyst*) and P_3 (*software engineer*). The edges weight are the implication relative frequency. The relative frequency was calculated according to equation: $\mathscr{F} = \frac{F_i}{F_p}$, where $\mathscr{F}$ is the relative frequency, F_i is the

[2]ABAP: Advanced Business Application Programming
[3]BI: Business Intelligence

implication absolute frequency and F_p is $|m'|$. For example: the implication $\{C\ language\} \rightarrow \{software\ engineer\}$ represents 31 objects (Fi) among 59 objects that have *software engineer* as incidence (Fp). So, this implication relative frequency is 0.52. For obtain the result as percentage is only multiply by 100, generating the relative frequency percentage of 52% to this proper implication in *software engineer* set of implications. Therefore, the thicker edges represents implications with greater relative frequency. It is important to note that, we applied the relative frequency measure, because in this case, the frequency represents the significance of an implication inside its class. And the local significance is more relevant than the implication support in the complete proper implications set. For example, the implication $\{java\ frameworks\} \rightarrow \{software\ engineer\}$ has relative frequency of 75.56%, and its support is 2.47%. In this example, to analyze the relative frequency is more important than the implication support, because the specific objective is identify the conditions to reach a *software engineer* job. Is important to note that, in both cases the confidence is 100%.

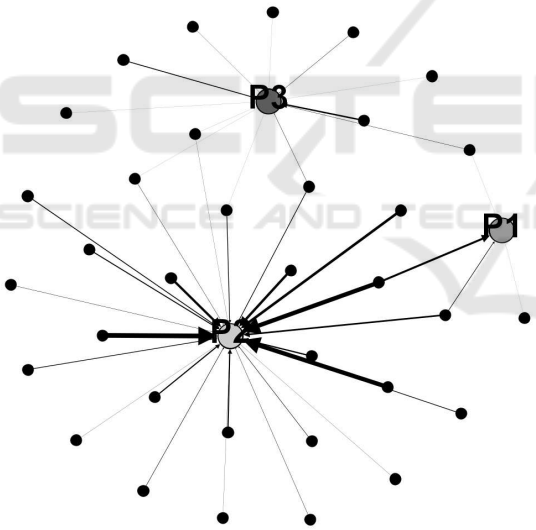

Figure 6: Top 3 jobs and their skills, where P_1 is *data scientist*, P_2 is *information security analyst* and P_3 is *software engineer* job position.

Figure 6 shows the intersections between the minimal sets of skills, considering the top 3 jobs positions described above. In this case, the nodes P_1 and P_3 share four set of skills like $\{BI\} \rightarrow \{security\ analyst\}$ and $\{BI\} \rightarrow \{software\ engineer\}$. The nodes P_1 and P_2 share two set of skills, like the proper implications show: $\{agile\ methodology\} \rightarrow \{data\ scientist\}$ and $\{agile\ methodology\} \rightarrow \{information\ security\ analyst\}$. And, the nodes P_1 and P_3 share only one set of skills, on $\{active\ directory^4\} \rightarrow \{data\ scientist\}$ and $\{active\ directory\} \rightarrow \{software\ engineer\}$.

From these intersections, we observed that the greater the intersections amount between skills sets, more similar are the requirements to achieve a position. It would indicated possibilities to professional mobility among positions, when the set of skills (premises) implies in several different positions (conclusions). So a professional could have competence to assume different positions, because his skills could be applied to different jobs. For example, in recruitment and selection hiring process, this professional could be compatible with several job vacancy, therefore he could be more jobs opportunities. Another example is the case when a professional needs change jobs, his skills allow greater career mobility.

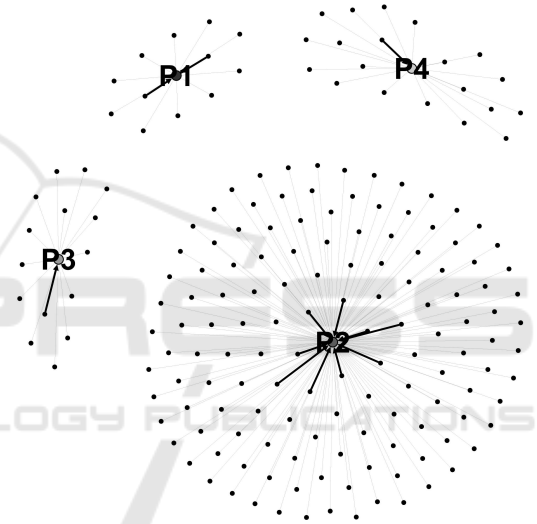

Figure 7: IT career hierarchical levels, where P_1 to P_4 represents the following job positions: P_1 is *IT analyst*, P_2 is *IT manager*, P_3 is *IT coordinator* and P_4 is *IT director*.

Figure 7 shows 4 positions that represents different hierarchical levels of IT career. The central nodes represent these 4 positions: P_1 (IT analyst), P_2 (IT coordinator), P_3 (IT manager) and P_4 (IT director). The other nodes represent minimal sets of skills, and edges represent implications. It is important to note that, edges weight was calculated using the relative frequency, previously described. From the figure we could observe that there are disjoint sets and there are not any intersections among positions. According to this hierarchy, P_1 and P_2 are positions related to the early career, while P_3 and P_4 are positions hierarchically superior. So, for P_1 was expected technical skills like in the proper implication {*.NET, automa-*

[4]Active directory: Microsoft tool kit for store and control information about network configurations.

tion systems$\} \rightarrow \{$*IT analyst*$\}$. P_2 involves skills that represent the transition between technical and managerial level, like in the proper implication $\{$*.NET, data base, ERP, it governance*$\} \rightarrow$ *IT coordinator*. P_3 also involves skills related to hierarchical transition, but it was expected more managerial than technical skills, it could be expressed by the implication $\{$*BPM*[5]*, cloud computing, CRM*[6]$\} \rightarrow$ *IT manager*. Finally, an IT director (P_4) have to develop managerial skills like was identified in implication $\{$*assets management, BI, business management, consulting*$\} \rightarrow$ *IT director*. Then, for the professional get a career advancement, he have to develop skills of different natures.

6 CONCLUSION AND FUTURE WORKS

In this paper, we presented an FCA-based approach to identify professional behavior through data scraped from *LinkedIn*. Specifically, we apply proper implications to identify the minimum sets of skills that is necessary for achieve a job position. In this case, firstly, we model the problem domain to understand the features which characterize a person as a professional, according to *model of competence*. After, we scraped data from *LinkedIn*, we applied preprocessing techniques and transform this data into a formal context. Finally, we extracted proper implications through *PropIm* algorithm and analyze the results through graph representations.

The main contributions of this paper were: a professionals data set scraped from *LinkedIn*, a FCA-based approach, and experiments set for apply FCA to professional career analysis.

As part of FCA-based approach, we propose the *PropIm* algorithm. The goal of *PropIm* algorithm is extract proper implications with support greater than zero. It was implemented applying pruning heuristics and scalable strategy. In future works the algorithm will be modified to run as a distributed application. The problem's order complexity to extract proper implications from formal context is $O(|M||\mathscr{I}|(|G||M| + |\mathscr{I}||M|))$. The proposed algorithm also has exponential complexity, but the implemented strategy reduce the computational effort computing only implications with support greater than zero, and the pruning heuristic reduce the possibilities of attributes combinations into premises. Moreover, the initial loop allows process the context in distributed way, because each conclusion from formal context can be processed separately without causing loss of information in the final set of proper implications.

The experiments answer the two 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?

In both questions, the proper implications was showed as graph representation and the analysis is based in central nodes density, intersections among premises and edges weight. For the first question, experiments show that denser nodes represent positions with more diversification of minimum sets of skills, while less dense nodes represent jobs that require more specific sets of skills. For the second question, experiments show that there are intersections between sets of skills from different jobs positions. These intersections means that same set of skills is required for different positions, which allows more professional opportunities in different industries and more professional mobility. We also analyzed a set of positions that compose a hierarchy of jobs in IT area. With this, it was observed that disjointed sets were formed without intersections between the skills, which shows that for a professional to progress of career, it is necessary to develop skills of different natures.

As future work, we intend to exploit other algorithms, particularly those capable of obtaining the set of implications from concept or the subset of formal concepts as proposed by Dias (2016). Moreover, we intend to implement the *PropIm* algorithm as a distributed application and compare several algorithms to extract proper implications from formal context. We also intend expand the experiments for all dimensions from the professional model of competence.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support received from the Foundation for Research Support of Minas Gerais state, FAPEMIG; the National Council for Scientific and Technological Development, CNPq; Coordination for the Improvement of Higher Education Personnel, CAPES. We would also express gratitude to the Federal Service of Data Processing, SERPRO.

[5]BPM: Business process management.

[6]CRM: Customer relationship management

REFERENCES

Ali, S. S., Bentayeb, F., Missaoui, R., and Boussaid, O. (2014). *An Efficient Method for Community Detection Based on Formal Concept Analysis*, pages 61–72. Springer International Publishing, Cham.

Atzmueller, M. (2015). Subgroup and community analytics on attributed graphs. In *Proceedings of the Workshop on Social Network Analysis using Formal Concept Analysis*.

Aufaure, M.-A. and Le Grand, B. (2013). Advances in fca-based applications for social networks analysis. *Int. J. Concept. Struct. Smart Appl.*, 1(1):73–89.

Banerjee, S., Badr, Y., and Al-shammari, E. T. (2014). *Social Networks: A Framework of Computational Intelligence*, volume 526. Springer Berlin Heidelberg.

Barysheva, A., Golubtsova, A., and Yavorskiy, R. (2015). Profiling less active users in online communities. In *Proceedings of the Workshop on Social Network Analysis using Formal Concept Analysis*.

Bertet, K. and Monjardet, B. (2010). The multiple facets of the canonical direct unit implicational basis. *Theoretical Computer Science*, 411(22–24):2155 – 2166.

Brandão, H. P. and Guimarães, T. d. A. (2001). Gestão de competências e gestão de desempenho: tecnologias distintas ou instrumentos de um mesmo construto? *Revista de Administração de empresas*, 41(1):8–15.

Cast, C. (2016). Jobs rated report 2016: Ranking 200 jobs. Accessed in 2016-12-12.

Codocedo, V., Baixeries, J., Kaytoue, M., and Napoli, A. (2016). Contributions to the formalization of order-like dependencies using fca. In *Proceedings of the 5th International Workshop What can FCA do for Artificial Intelligence*. CEUR-WS.

Cordero, P., Enciso, M., Mora, A., Ojeda-Aciego, M., and Rossi, C. (2015). Knowledge discovery in social networks by using a logic-based treatment of implications. *Know.-Based Syst.*, 87(C):16–25.

Cuvelier, E. and Aufaure, M.-A. (2011). A buzz and e-reputation monitoring tool for twitter based on galois lattices. In *Conceptual Structures for Discovering Knowledge*, pages 91–103. Springer, Berlin Heidelberg.

Dias, S. M. (2016). *Redução de Reticulados Conceituais (Concept Lattice Reduction)*. PhD thesis, Department of Computer Science of Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil. In Portuguese.

Durand, T. (1998). Forms of incompetence. In *Proceedings Fourth International Conference on Competence-Based Management. Oslo: Norwegian School of Management*.

Ganter, B., Stumme, G., and Wille, R. (2005). *Formal concept analysis: foundations and applications*, volume 3626. Springer Science & Business Media.

Ganter, B. and Wille, R. (2012). *Formal concept analysis: mathematical foundations*. Springer Science & Business Media.

Jota Resende, G., De Moraes, N. R., Dias, S. M., Marques Neto, H. T., and Zarate, L. E. (2015). Canonical computational models based on formal concept analysis for social network analysis and representation. In *Web Services (ICWS), 2015 IEEE International Conference on*, pages 717–720. IEEE.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40(10):4065 – 4074.

Krajči, S. (2014). *Social Network and Formal Concept Analysis*, pages 41–61. Springer International Publishing, Cham.

Li, L., Zheng, G., Peltsverger, S., and Zhang, C. (2016). Career trajectory analysis of information technology alumni: A linkedin perspective. In *Proceedings of the 17th Annual Conference on Information Technology Education*, SIGITE '16, pages 2–6, New York, NY, USA. ACM.

LinkedIn (2016). About linkedin. Accessed in 2016-12-02.

Lorenzo, E. R., Cordero, P., Enciso, M., Missaoui, R., and Mora, A. (2016). Caisl: Simplification logic for conditional attribute implications. In *CLA*.

Neto, S. M., Song, M., Dias, S., et al. (2015a). Minimal cover of implication rules to represent two mode networks. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 211–218. IEEE.

Neto, S. M., Song, M. A., Dias, S. M., and Zárate, L. E. (2015b). Using implications from fca to represent a two mode network data. *nternational Journal of Software Engineering and Knowledge Engineering (IJSEKE)*.

Neznanov, A. and Parinov, A. (2015). Analyzing social networks services using formal concept analysis research toolbox. In *Proceedings of the Workshop on Social Network Analysis using Formal Concept Analysis*.

Rome, J. E. and Haralick, R. M. (2005). *Towards a Formal Concept Analysis Approach to Exploring Communities on the World Wide Web*, pages 33–48. Springer Berlin Heidelberg, Berlin, Heidelberg.

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. " O'Reilly Media, Inc.".

Snasel, V., Horak, Z., Kocibova, J., and Abraham, A. (2009). Analyzing social networks using fca: Complexity aspects. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM International Joint Conferences on*, volume 3, pages 38–41.

Soldano, H., Santini, G., and Bouthinon, D. (2015). Abstract and local concepts in attributed networks. In *Proceedings of the Workshop on Social Network Analysis using Formal Concept Analysis*.

Stattner, E. and Collard, M. (2012). Social-based conceptual links: Conceptual analysis applied to social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 25–29.

Taouil, R. and Bastide, Y. (2001). Computing Proper Implications. In *Proceedings of the International Confer-*

ence on Conceptual Structures - ICCS, pages 46–61, Stanford, CA US.

Xu, Y., Li, Z., Gupta, A., Bugdayci, A., and Bhasin, A. (2014a). Modeling professional similarity by mining professional career trajectories. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1945–1954. ACM.

Xu, Y., Li, Z., Gupta, A., Bugdayci, A., and Bhasin, A. (2014b). Modeling professional similarity by mining professional career trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1945–1954. ACM.