Expert Systems with Applications 42 (2015) 7084–7097

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Concept lattices reduction: Definition, analysis and classification

Sérgio M. Dias [a,b,*], Newton J. Vieira [a]

[a] Department of Computer Science, Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, ICEx, 4010, Pampulha, 31.270-901 Belo Horizonte, Minas Gerais, Brazil
[b] Federal Service of Data Processing (SERPRO), Av. José Cândido da Silveira, 1.200, Cidade Nova, 31.035-536 Belo Horizonte, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Available online 9 May 2015

Keywords:
Formal concept analysis
Concept lattices
Reduction

ABSTRACT

Formal concept analysis (FCA) is currently considered an important formalism for knowledge representation, extraction and analysis with applications in different areas. A problem identified in several applications is the computational cost due to the large number of formal concepts generated. Even when that number is not very large, the essential aspects, those effectively needed, can be immersed in a maze of irrelevant details. In fact, the problem of obtaining a concept lattice of appropriate complexity and size is one of the most important problems of FCA. In literature, several different approaches to control the complexity and size of a concept lattice have been described, but so far they have not been properly analyzed, compared and classified. We propose the classification of techniques for concept lattice reduction in three groups: *redundant information removal*, *simplification*, and *selection*. The main techniques to reduce concept lattice are analyzed and classified based on seven dimensions, each one composed of a set of characteristics. Considerations are made about the applicability and computational complexity of approaches of different classes.

1. Introduction

Formal concept analysis (FCA) is currently considered an important formalism for knowledge representation, extraction and analysis. Its formalization was born in 1982 with the work of Wille (1982), who proposed considering each lattice element as a formal concept and the lattice itself as representing a conceptual hierarchy (Ganter & Wille, 1999).

The initial data in FCA is a binary relation $I \subseteq G \times M$, where the elements of G are called *objects* and the elements of M are called *attributes*. The triple (G, M, I) is called a *formal context*. From such a formal context, *formal concepts* are obtained, which are ordered pairs (A, B), where $A \subseteq G$ (the concept *extension*), $B \subseteq M$ (its *intention*) and each object in A has all the attributes in B and each attribute in B is an attribute common to all objects in A. The great potential of FCA is propitiated by the *concept lattice*, a complete lattice formed by the set of formal concepts extracted from the formal context. In fact, the main applications make use of the concept lattice, usually represented by means of a line diagram, or a nested line diagram, or a tree diagram, etc.

Even a small set of data can result in a very large number of formal concepts (Belohlavek & Macko, 2011). In fact, FCA induces potentially high combinatorial complexity and the structures obtained, even from a small dataset, may become prohibitively large (Klimushkin, Obiedkov, & Roth, 2010). Despite the fact that the worst case ($2^{min(|G|,|M|)}$) is rarely found in practice (Godin, Saunders, & Gecsei, 1986), the computational cost is still too prohibitive for many applications. Furthermore, the resulting number of formal concepts and the complexity of the relationships between concepts can make the analysis of the final lattice difficult (Rice & Siff, 2001). In particular, key aspects, those which are effectively sought, can be immersed in a maze of irrelevant details. Regardless of the number of formal concepts generated in the worst case, all relationships between concepts are present in the concept lattice. This feature is suitable in terms of completeness, but generally results in a large number of relationships, thus overloading the lattice's structure.

In fact, the problem of obtaining a concept lattice of appropriate size and structure, one that exposes the truly relevant aspects, is one of the most important problems of using FCA (Belohlavek & Macko, 2011; Belohlavek, Sklenár, & Zacpal, 2004a; Belohlavek & Vychodil, 2006; Klimushkin et al., 2010; Kuznetsov, Obiedkov, & Roth, 2007; Li, Mei, & Lv, 2012; Wei, Qi, & Zhang, 2008; Medina, 2012; Pernelle, Rousset, Soldano, & Ventos, 2002; Rice & Siff, 2001; Snásel, Polovincak, Abdulla, & Horak, 2008; Soldano, Ventos, Champesme, & Forge, 2010; Stumme, Taouil, Bastide, Pasquier, & Lakhal, 2002; Ventos & Soldano, 2005; Zhang, Wei, & Qi, 2005). Notice that it is an instance of the more general problem

* Corresponding author.
E-mail addresses: mariano@dcc.ufmg.br, sergio.dias@serpro.gov.br (S.M. Dias), nvieira@dcc.ufmg.br (N.J. Vieira).

http://dx.doi.org/10.1016/j.eswa.2015.04.044
0957-4174/

of looking for patterns, which occurs in many methods of data analysis and that frequently deal with the generation of an excessive number of patterns (Belohlavek & Macko, 2011).

There are many techniques for concept lattice reduction, each with different characteristics. Some of the techniques remove redundant information from the concept lattice. In general, they aim to find the minimum set of objects or attributes that keep the structure of the original concept lattice unchanged (Li, Mei, & Lv, 2011a; Medina, 2012; Pei & Mi, 2011; Shao, 2005; Wang & Ma, 2006; Wang & Zhang, 2008a; Zhang et al., 2005). Other techniques try to construct an abstraction of the concept lattice; in other words, they seek to obtain a high-level of simplification that exposes the truly important aspects (Belohlavek & Vychodil, 2009; Cheung & Vogel, 2005; Codocedo, Taramasco, & Astudill, 2011; Dias & Vieira, 2010; Gajdos, Moravec, & Snásel, 2004; Kumar & Srinivas, 2010b; Poshyvanyk & Marcus, 2007; Snásel, Dahwa Abdulla, & Polovincak, 2007). Finally, a third class of techniques works by selecting formal concepts, objects or attributes through a relevance criterion (Arévalo, Berry, Huchard, Perrot, & Sigayret, 2007; Godin & Mili, 1993; Kuznetsov, 1990; Riadh, Le Grand, Aufaure, & Soto, 2009; Rice & Siff, 2001; Stumme et al., 2002). Different characteristics are observed in each of the three classes. However, so far, those characteristics have not been properly identified and used to analyze and compare the existing classes.

In this paper, the three classes of techniques mentioned above are elaborated further, and the main existing techniques of each class are identified. Formal concept analysis is used to analyze the techniques based on seven dimensions, each one consisting of a set of characteristics. Along with the FCA based analysis, considerations are carried out about computational complexity, feasibility and quality of the resulting concept lattice.

The remaining sections of this paper are organized as follows. Section 2 reviews the core notions and terminology of FCA, shows a small illustrative example, and defines techniques for concept lattice reduction. Section 3 classifies the existing reduction techniques into three main classes. Section 4 identifies seven dimensions for the analysis of the reduction techniques. Section 5 summarizes the main techniques for concept lattice reduction and performs an analysis of those techniques by means of FCA. Related works are identified in Section 6. Finally, conclusions are presented in Section 7.

2. FCA: core notions and a small illustrative example

This short review presents the notions and terminology which are important for the understanding of our work, and introduces an example to be used throughout the text. The notions and terminology are based on the excellent book of Ganter and Wille (1999).

As said in the introduction, in FCA the initial data are presented as a *formal context*, a triplet (G, M, I), where G is a set of elements called *objects*, M is a set of elements called *attributes*, and $I \subseteq G \times M$ is called an *incidence relation*. If $(g, m) \in I$, one says that "*the object g has the attribute m*". A formal context is usually presented as a cross table where the objects are row headers, the attributes are column headers, and there is a cross in row g and column m if and only if $(g, m) \in I$. Table 1 shows an example of formal context.

Given a set of objects $A \subseteq G$ from a formal context (G, M, I), the set of attributes which are common to all those objects is termed A'. Similarly, for a set $B \subseteq M$, B' is the set of objects that have all the attributes from B. That is to say $A' = \{m \in M | \forall g \in A (g, m) \in I\}$ and $B' = \{g \in G | \forall m \in B (g, m) \in I\}$.[1] By using such *derivation operators*, the notion of *formal concept* is defined as a pair $(A, B) \in \mathcal{P}(G) \times \mathcal{P}(M)$ such that $A' = B$ and $B' = A$, where A is called the *extent* and B the *intent* of the concept. For example, from the formal context of Table 1, it can be seen that the pair

$$(\{3, 4\}, \{b, d, f\})$$

is a formal concept with extent $\{3, 4\}$ and intent $\{b, d, f\}$.

Table 1
A formal context.

Obj/Att	*a*	*b*	*c*	*d*	*e*	*f*
1	×	×				
2	×					
3	×	×	×	×		×
4		×		×	×	×
5	×					

The set of formal concepts is ordered by the partial order $\preceq$ such that for any two formal concepts (A_1, B_1) and (A_2, B_2), $(A_1, B_1) \preceq (A_2, B_2)$ iff $A_1 \subseteq A_2$ (equivalently, $B_2 \subseteq B_1$). The set of concepts ordered by $\preceq$ constitutes a complete lattice (Davey & Priestley, 1990), the so called *concept lattice*. The concept lattice obtained from a formal context (G, M, I) is denoted $\mathcal{B}(G, M, I)$. Fig. 1 presents the line diagram (Ganter & Wille, 1999) of the concept lattice originated from the formal context of Table 1.[2] Each node in the line diagram represents a formal concept. The objects are shown inside white boxes drawn below some concept nodes and the attributes inside gray boxes drawn above some concept nodes. The boxes are distributed in such a way that the extent of a concept is obtainable by collecting all objects from the concept node to the lattice infimum, and its intent is obtainable by visiting all attributes from the concept node to the lattice supremum. The labeling can also be explained by means of the notions of *object concept* and *attribute concept*. Given an object $g, \gamma g$ is the object concept (g'', g'), and μm, for an attribute m, is the attribute concept (m', m''). Then, the labeling of $\mathcal{B}(G, M, I)$ proceeds as follow: for each object g the formal concept γg is labeled g and for attribute m the formal concept μm is labeled m.

The first part of the *basic theorem* on concept lattices (Wille, 1982) says that a concept lattice $\mathcal{B}(G, M, I)$ is a complete lattice in which for any arbitrary set $C \subseteq \mathcal{B}(G, M, I)$ the *infimum* and *supremum* are given by

$$\bigwedge C = \left(\bigcap X, \left(\bigcup Y\right)''\right) \quad \text{and} \quad \bigvee C = \left(\left(\bigcup X\right)'', \bigcap Y\right)$$

where $X = \{A \,|\, (A, B) \in C\}$ and $Y = \{B \,|\, (A, B) \in C\}$.

The knowledge embodied by a formal context (G, M, I) or its concept lattice $\mathcal{B}(G, M, I)$ can be used to derive implications $P \to Q$, where P and Q are sets of attributes, which express that $P' \subseteq Q'$, or in other words, if an object has all the attributes in P, then it has all those in Q (equivalently, $Q \subseteq P'$) (Ganter & Wille, 1999). Notation: the set brackets will be omitted in both sets of an implication. For example, from the lattice of Fig. 1 we have the implication $a, d \to c$, meaning that those objects who have a and d as attributes (actually, only the object 3) also have c. If X is a set of attributes, then X *respects* an implication $P \to Q$ iff $P \neg \subseteq X$ or $Q \subseteq X$. An implication $P \to Q$ *holds* in a set $\{X_1, \ldots, X_n\} \subseteq \mathcal{P}(M)$ 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. $P \to Q$ *follows* from a set of implications $\mathcal{I}$ iff for every set of attributes X if X respects $\mathcal{I}$, then it respects $P \to Q$. A set of implications $\mathcal{I}$ is said to be *complete* iff every implication of (G, M, I) follows from $\mathcal{I}$. Of particular importance are non-redundant sets of implications. A set of implications

[1] The notation x' will be used as abbreviating $\{x\}'$, whether x is an object or an attribute.

[2] All line diagrams in this paper were drawn using the Conexp software (Yevtushenko, 2000).