Expert Systems With Applications 44 (2016) 198–216

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Supporting scientific knowledge discovery with extended, generalized Formal Concept Analysis

Francisco J. Valverde-Albacete [a], José María González-Calabozo [b], Anselmo Peñas [a], Carmen Peláez-Moreno [b,*]

[a] Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid 28040, Spain
[b] Departamento de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, Leganés 28911, Spain

A R T I C L E I N F O

Keywords:
Scientific knowledge discovery
Exploratory Data Analysis
Landscapes of Knowledge
Metaphor theory
Formal Concept Analysis
$\mathcal{K}$-Formal Concept Analysis
Extended Formal Concept Analysis
Semiring theory
Confusion matrix
Relation extraction
Gene expression data

A B S T R A C T

In this paper we fuse together the Landscapes of Knowledge of Wille's and Exploratory Data Analysis by leveraging Formal Concept Analysis (FCA) to support data-induced scientific enquiry and discovery.

We use extended FCA first by allowing $\mathcal{K}$-valued entries in the incidence to accommodate other, non-binary types of data, and second with different modes of creating formal concepts to accommodate diverse conceptualizing phenomena.

With these extensions we demonstrate the versatility of the Landscapes of Knowledge metaphor to help in creating new scientific and engineering knowledge by providing several successful use cases of our techniques that support scientific hypothesis-making and discovery in a range of domains: semiring theory, perceptual studies, natural language semantics, and gene expression data analysis.

While doing so, we also capture the affordances that justify the use of FCA and its extensions in scientific discovery.

1. Introduction

Formal Concept Analysis (FCA) can be understood as an unsupervised analysis technique for matrices with boolean entries (Ganter & Wille, 1999). The underlying theoretical construct is a Galois Connection induced by a relation $I \subseteq 2^{G \times M}$ between a set of objects G and a set of attributes M[1]. In this case, the Galois connections adopts the guise of a pair of functions, the *polars*, induced by the relation transforming sets of objects into sets of attributes and vice versa: pairs of sets of objects and attributes that are mutually transformed are called *formal concepts* and the inclusion order of object sets (dually, that of attribute sets) turns out to be a complete lattice, providing an excellent canvas against which to cast many of the properties hidden in the data. A lightweight introduction to FCA can be found in Section 2.1.

In spite of the width of the work trying to make FCA a sound technique for the exploratory analysis of knowledge, as reviewed, for instance in Poelmans, Kuznetsov, Ignatov, and Dedene (2013b), Poelmans, Ignatov, Kuznetsov, and Dedene (2013a), it has found but a feeble echo in the Statistical and Machine Learning communities.

Tukey was the figure who crystallized the cry for exploratory, discovery-driven methods in the Statistics community. He called his proposal *Exploratory Data Analysis* (EDA) and advocated a complementary curriculum for Statisticians balancing EDA against Statistical Hypothesis Testing (that he called Confirmatory Data Analysis) (Tukey, 1980). At present, EDA is standard practice of good statistics, it is taught at basic statistics courses in academia (Tukey, 1977), and supported by widely-used data processing software (Matlab, 2012; R Core Team, 2014).

FCA falls under the definition of an EDA technique for boolean matrices, yet very few work in the FCA community or otherwise, highlights this fact. Furthermore, extensions to FCA exist that allow it to work with more quantitative data like FCA in a fuzzy setting (Bělohlávek, 2002) or $\mathcal{K}$-FCA (Valverde-Albacete & Peláez-Moreno, 2006) where the numeric data take values in different kinds of semirings.

1.1. (Formal Conceptual) Landscapes of Knowledge

One of the founders of FCA, Wille, a mathematician and philosopher, developed in the late 1990s a program for knowledge

* Corresponding author. Tel.: +34 916248771; fax: +34 916248749.
E-mail addresses: fva@lsi.uned.es (F.J. Valverde-Albacete), melopsitaco@gmail.com (J.M. González-Calabozo), anselmo@lsi.uned.es (A. Peñas), carmen@tsc.uc3m.es (C. Peláez-Moreno).

[1] G for German *Gegenstunde* "object" and M for German *Merkmale* "attribute" are customary in FCA theory.

http://dx.doi.org/10.1016/j.eswa.2015.09.022
0957-4174/

Table 1
The activities supporting the Landscape of Knowledge metaphor, their descriptions (from (Wille, 1999)) and some aliases that we use for them. The last two rows are our proposal for inclusion in the list.

Activity	Description	Aliases
Focusing	Selecting what one wants to look into.	*contextualizing*
Exploring	"...looking for something of which one only has a vague idea..."	*browsing*
Searching	"...looking for something which one can more or less specify but not localize..."	
Recognizing	"...perceiving clearly circumstances and relationships..."	*Clarifying*
Identifying	"...determining the taxonomic position of an object within a given classification..."	*categorizing, placing in a taxonomy*
Analyzing	"...examining data in their relationships while guided by the theoretical views and declared purposes..."	
Investigating	"...study by close examination and systematic enquiry..."	
Deciding	"...resolving a situation of uncertainty by an order..."[a]	
Improving	"...enhancement in quality and value. ..."	*distilling, valorizing*
Restructuring	"...to reshape a given structure, which, within the scope of our discussion, is conceptual in its nature..."	
Memorizing	"...a process of committing and reproducing what has been learned and retained. ..."	*committing to memory*
Hypothesizing	Abducing facts and relationships based on data.	*forming an opinion*
Indexing	Indexing other knowledge resources using concepts, objects or attributes	*interfacing resources*

[a] There is a possible ambiguity here in the meaning of *order*, in spite of Wille's being an *order mathematician* we have decided to understand here *command*.

discovery based in FCA. He baptized it the "Conceptual Landscapes of Knowledge Paradigm" (LofK) but we believe it can be better understood as a *metaphor* in the spirit of Lakoff and Johnson (1996).

Metaphorical interpretation eases the comprehension of an unfamiliar domain of knowledge in terms of an already-known domain: the entities and issues in the new domain are mapped onto the well-known domain to make them more apprehensible. For a review of tenets and notation in metaphor theory see Lakoff (1987), Lakoff and Johnson (1996).

The basic metaphor in LofK seems to be

Metaphor 1. UNDISCOVERED KNOWLEDGE IS UNCHARTED TERRITORY.

and is founded in the analogy between the investigation of hitherto unknown knowledge and the physical exploration of non-mapped terrain. The analogy is reinforced by a series of subsidiary metaphors like

Submetaphor 2. THE INTRICACIES OF KNOWLEDGE ARE FEATURES OF THE LANDSCAPE.

This emphasizes that navigating the nitty-gritty details of knowledge is like exploring difficult to access places.

Once the initial identification of knowledge and terrain is accepted, Wille proposes that FCA can help in fleshing out this metaphor by providing a physical embodiment for knowledge that can be acted upon as if it were a *map of the terrain*. The first step of this exploration is to explicitly highlight the data being considered by forming a *formal context* which is a two-mode relation between objects and attributes. This action is called *contextualizing*:

Metaphor 3. BUILDING A FORMAL CONTEXT IS SETTING YOURSELF ON A VANTAGE POINT ON THE KNOWLEDGE LANDSCAPE.

At the same time, contextualizing sets the limits of the knowledge you can see, since you choose to concentrate on some information to the detriment of some other. This is further emphasized by the context-lattice duality of FCA and the following

Metaphor 4. A CONCEPT LATTICE IS A MAP OF THE KNOWLEDGE LANDSCAPE.

The different operations on the formal context would subsequently represent acting on the territory being charted, with the idea of using the concept lattice to explore the unknown territory until it becomes tame.

Table 1 lists the knowledge-exploring activities supported by these metaphors. Most of these activities are expressed in either the main or the secondary domain of the metaphor, like "browsing" in the physical domain, and "analyzing" in the cognitive domain. Note that these metaphors have already been put to the test in the building of semantic file systems (Martin, 2004; Martin & Eklund, 2005) and Logical Information Systems (Ferré, 2007; Ferré & Ridoux, 2001) as Information Retrieval systems. A thorough review of the use of FCA to model knowledge is Poelmans et al. (2013b), while applications are detailed in Poelmans et al. (2013a).

In later work, Wille has concentrated in developing these metaphors with a view towards making them *actionable* by providing epistemic interpretation for the constructs of FCA (Wille, 2006). But even his early papers show usage of the actions enunciated in Table 1 or precursors of them.

1.2. FCA as an instance of Exploratory Data Analysis

We believe that the following metaphor captures one of the tenets of exploratory analysis:

Metaphor 5. KNOWLEDGE IS AN EXHIBIT

We refer here to "exhibit" much as in a work of art in a museum or an item in a shop. The idea is that knowledge can be taken "out in the open" to be shared, subjected to public scrutiny and commentary.

In our experience, FCA is a method of EDA (Tukey, 1977) that supports the knowledge-as-exhibit metaphor through two submetaphors:

Submetaphor 6. KNOWLEDGE CAN BE VISUALIZED.

This is achieved through the Hasse diagram of the concept lattice, while

Submetaphor 7. KNOWLEDGE CAN BE (BODILY) MANIPULATED.

is achieved through (a) the polars that enable the transformation of objects and attributes into formal concepts, and (b) the operations between formal concepts allowing us to obtain more general or more specific ones.

A summary of metaphors to be fleshed out can be consulted in Table 2.

The first endeavor of this paper is to show how FCA as interpreted in the LofK metaphors can be leveraged as an EDA technique for scientific and engineering discovery by helping with the elicitation, structuring and manipulation of knowledge. The first result of this endeavor in this paper is *a methodology to carry out exploratory analysis of two-mode data with (extended, generalized) FCA*, which we put to the test and exemplify in Section 3.

1.3. Extensions to Formal Concept Analysis

Unfortunately, the fact that *standard* FCA can only deal with boolean incidences often prevents scientists with other type of data