Logical Scaling in Formal Concept Analysis

Susanne Prediger

Technische Hochschule Darmstadt, Fachbereich Mathematik
Schloßgartenstr. 7, D–64289 Darmstadt, prediger@mathematik.th-darmstadt.de

Abstract. Logical scaling is a new method to transform data matrices which are based on object-attribute-value-relationships into data matrices from which conceptual hierarchies can be explored. The derivation of concept lattices is determined by terminologies expressed in a formal-logical language.

Published in: D. Lukose et.al. (eds.): Conceptual Structures: Fulfilling Peirce's Dream. Proceedings of the ICCS'97, LNAI 1257, Springer, Berlin 1997, 332–341.

1 Introduction

The aim of formal concept analysis is to explore conceptual patterns in empirical data contexts. Methods have been developed to find conceptual hierachies and to represent them in line diagrams based on concept lattices (cf. [Wi82], [GW96]). These methods can be of great interest for knowledge representation and data mining. Concept lattices can also be relevant as principled ways to structure the type lattices used for conceptual graphs.

In general, there is no immediate, "automatic" way to derive the conceptual structures of data contexts which are based on object-attribute-value relationships. The first approach to transform these data contexts into concept lattices, namely *conceptual scaling*, was presented in [GSW86]. Since then, a great variety of applications in formal concept analysis has been based on this method.

In this paper, conceptual scaling will be rediscussed and compared with a new, alternative method called *logical scaling* in which the derivation of concept lattices is determined by terminologies expressed in a formal-logical language.

2 Contexts derived by conceptual scales

Object-attribute-value-relationships are a frequently used data structure to code real-world problems. In formal concept analysis, they are formalized in *many-valued contexts*. (Note that the word "context" is used in a special way here. It has nothing to do with the contexts of Sowa's conceptual structures (cf. e. g. [So92])).

In [GW89], a *many-valued context* is formally defined as a quadruple (G, M, W, I), where G, M, and W are sets whose elements are called *objects, (many-valued) attributes* and *attribute values* respectively, and I is a ternary relation with $I \subseteq G \times M \times W$ such that $(g, m, v) \in I$ and $(g, m, w) \in I$ always

$\mathbb{K}$	importance	sports	books	home	events
1	5	su	$\{g, e\}$	am	sp
2	3	sa	$\{e\}$	ki	cu
3	1	wa	$\{s, g\}$	am	cu
4	2	tf	$\{s, e\}$	fs	po
5	4	sk	$\{g, e\}$	hg	sp
6	2	hm	$\{s, g, e\}$	am	cu
7	1	wa	$\{s, g, e\}$	am	cu
8	3	sk	$\{g, e\}$	ki	po
9	4	sk	$\{g, e\}$	hg	sp
10	5	su	$\{e\}$	fs	sp
11	2	jo	$\{s, g\}$	fs	cu
12	1	wa	$\{s, g, e\}$	am	cu
13	4	su	$\{g\}$	hg	sp
14	5	sa	$\{e\}$	hg	sp
15	3	jo	$\{s, e\}$	ki	cu

Fig. 1. Many-valued context *leisure activities*

implies $v = w$. An attribute m of a many-valued context (G, M, W, I) may be considered as a partial map of G into W which suggests to write $m(g) = w$ rather than $(g, m, w) \in I$. The context (G, M, W, I) is called n-valued if W has cardinality n. One-valued contexts correspond to formal contexts as defined in [Wi82].

For illustration, we recall one of the first examples of scaled many-valued contexts which was presented in [GSW86]. It deals with a data set extracted from a case study on leisure activities (see [AL83]). 15 persons are listed with attribute values concerning five many-valued attributes:

importance of leisure: (importance)	very important (5), rather important (4), important (3), less important (2), unimportant (1);
sports activities: (sports)	sailing (sa), surfing (su), skiing (sk), walking (wa), hiking in the mountains (hm), jogging (jo), track and field (tf);
book reading: (books)	all combinations of specific text books (s), general books (g) and entertaining books (e);
hobbies at home: (home)	house and garden (hg), kitchen (ki), art and music (am), family and social life (fs);
visit of events: (events)	culture (cu), sports (sp), politics (po).

The data set is represented by the many-valued context shown in figure 1.

The aim of every *scaling process* in formal concept analysis is to obtain a derived formal context (G, N, J) with the same objects as (G, M, W, I) whose extents can be thought of as the "meaningful" subsets of G. For this derived context, we can obtain a concept lattice as usual (cf. [Wi92] for an introduction). The concept lattice of the derived context is considered to be the conceptual structure of the many-valued context.

"Meaningful" refers to the interpretation of the data which can only be given by an expert of the field the data is from and never by the mathematician alone. This interpretation must always be purpose-oriented and should be founded on theoretical considerations.

The basic idea of *conceptual scaling* is to derive the context by conceptual scales. We assign to each many-valued attribute $m \in M$ a *conceptual scale* $\mathbb{S}_m$ which is a formal (one-valued) context $\mathbb{S}_m := (G_m, M_m, I_m)$ with $m(G) \subseteq G_m$. The choice of these scales is a matter of interpretation. The task is to select $\mathbb{S}_m$ in such a way that it reflects the implicitly given structure of the attribute values as well as the issues of data analysis.

The second step of conceptual scaling is to decide how the different many-valued attributes can be combined to describe concepts. The disjoint union of attribute sets often proves sufficient and we can thus restrict ourselves to looking at this so called *plain conceptual scaling*.

The many-valued context (G, M, W, I), together with the family of scales $(\mathbb{S}_m)_{m \in M}$ is called the *plainly scaled context* and determines the *derived context* which is defined as

$$(G, \bigcup_{m \in M} \{m\} \times M_m, J) \quad \text{where} \quad g\, J\, (m, w) \, :\Longleftrightarrow \, m(g)\, I_m\, w.$$

In our example, we extract the information about the implicit structures of the attribute values from [AL83] and scale the attribute importance and books

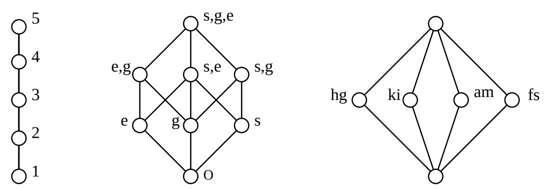

Fig. 2. Concept lattices of the scales $\mathbb{S}_{\text{importance}}$, $\mathbb{S}_{\text{books}}$ and $\mathbb{S}_{\text{home}}$

ordinally and the attribute home nominally.

$$\mathbb{S}_{\text{importance}} := (\{1,2,3,4,5\},\{1,2,3,4,5\},\leq),$$
$$\mathbb{S}_{\text{books}} := (\mathfrak{P}(\{\mathfrak{s},\mathfrak{g},\mathfrak{e}\}),\mathfrak{P}(\{\mathfrak{s},\mathfrak{g},\mathfrak{e}\}),\subseteq) \qquad \text{and}$$
$$\mathbb{S}_{\text{home}} := (W, W, =).$$

These scales can be represented by the concept lattices shown in figure 2. Assuming that politics is more similar to culture than to sports, the attribute events is scaled interordinally. The attribute values of sports are understood as structured by a tree-like hierarchy. In this way, we obtain two scales which are represented by the concept lattices in figure 3.

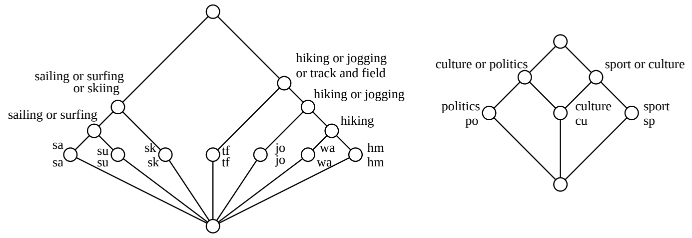

Fig. 3. Concept lattices of the scales $\mathbb{S}_{\text{sports}}$ and $\mathbb{S}_{\text{events}}$

In figure 4, we can see the concept lattice of the corresponding derived context where we can find interesting patterns. For example, none of the persons visiting sporting events reads specific books whereas all culturally interested persons do. More complex results are discussed in [GSW86].

Summarizing, we can say that conceptual scaling yields a global view of the conceptual patterns of data stored in many-valued contexts by applying expert knowledge about the inherent structure of some or all attribute values.
Sometimes, however, it is not the global view that is the desired result but the answer to more specific questions. In this case, we already have a certain conception of relevant combinations of attributes, so it is not necessary to scale all attributes. Instead, we use these relevant combinations of attributes to specify a limited terminology by which we can derive the context. This method is called *logical scaling.*

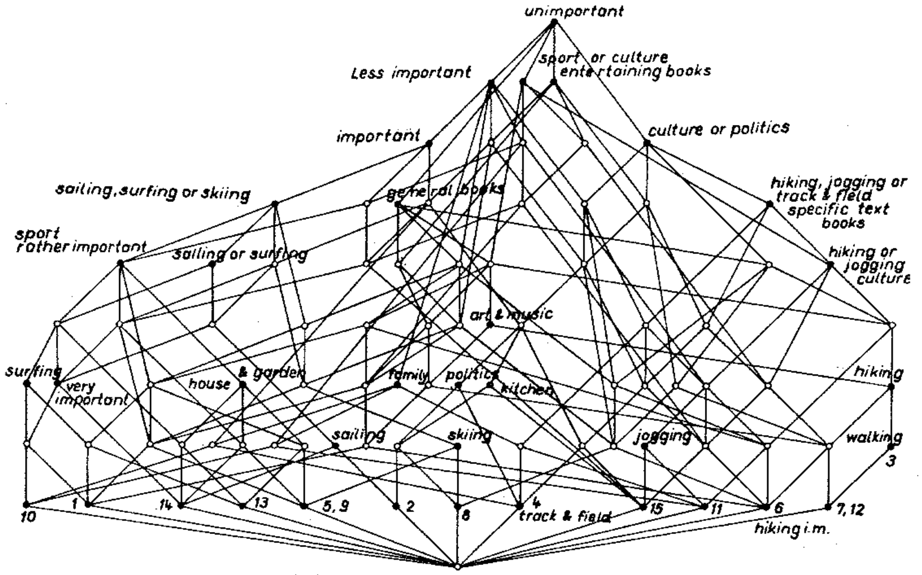

Fig. 4. Concept lattices of the derived context *leisure activities*

3 Contexts derived by terminologies

The basic idea of logical scaling consists of using a formal language to generate unary predicates from the attributes and attribute values of the many-valued context. These predicates form a terminology which determines a derived (one-valued) context.

For example, market researchers are interested in the leisure activities of certain groups of persons which are described by stereotypes of their behavioural patterns. They often examine groups and their stereotypes which are themselves results of data analysis like in [Fo77]. In our example, we define the relevant stereotypes by attributes and attribute values given in the many-valued context *leisure activities*. Then, we analyze how important leisure is for the specified groups. Therefore, the degrees of importance as well as the stereotypes must be described by unary predicates.

The formal language used in this article is based on elements of SQL (structured query language). This is a formal language which is utilized for queries and the administration of relational database management systems as Microsoft Access for example. Using SQL, one can enter the command

leisure unimportant:	(importance ≤ 1)
leisure less important:	(importance ≤ 2)
leisure important:	(importance ≤ 3)
leisure rather important:	(importance ≤ 4)
leisure very important:	(importance ≤ 5)
FAMILY MAN :	(home $=$ fs)
SPORTS FAN :	(events $=$ sp)
INTELLECTUAL:	(events $=$ po $\vee$ cu) $\wedge$ (books $\ni s$)
PRACTITIONER:	(events $=$ sp $\wedge$ home $\neq$ am) $\vee$ (books $= \{$e$\}$)

Fig. 5. Terminology *stereotypes*

`SELECT` Person `FROM` *activities* `WHERE` (events = cu `AND` home = am)

which asks for all persons in the data context *leisure activities* who prefer visiting cultural events and are interested in art and music at home. Here, we write the corresponding predicate in the following way:

$$(\mathsf{events} = \mathrm{cu} \wedge \mathsf{home} = \mathrm{am}).$$

As with many programming languages, SQL is quite intuitive so I refrain from explaining syntax and semantics here (cf. [In91] for details).

For our example, we use SQL to construct the terminology shown in figure 5 where four stereotypes and five degrees of importance of leisure are described. Formally, we define a *terminology* as a tuple $(\mathcal{P}, \mathcal{N}, \nu) =: \mathcal{T}$ where $\mathcal{P}$ is a set of unary predicates, N a set of names of attributes and ν a surjective *naming function* $\nu \colon N \to \mathcal{P}$.

When we *derive* the many-valued context $\mathbb{K} := (G, M, W, I)$ by the terminology $\mathcal{T} := (\mathcal{P}, \widetilde{\mathcal{M}}, \nu)$, we obtain the one-valued context $\mathbb{K}^{\mathcal{T}} := (G, N, E)$ where the relation E is given by the semantics of the formal language: for all $g \in G$ and $n \in N$ it holds

$$g\,E\,n :\Longleftrightarrow g \text{ satisfies } \nu(n).$$

This derivation can be conducted automatically in every database system using SQL as its query language whereas the choice of the terminology is determined by the research questions.

Figure 6 shows the context derived by our terminology *stereotypes* and the corresponding concept lattice. Looking at the concept lattice, we find interesting information about the dependencies between stereotypes and importance of leisure. For all seven INTELLECTUALS, leisure is not important or less important whereas for the sports fans leisure is much more important. All persons belonging to the stereotype SPORTS FAN consider leisure to be rather or even very important. The FAMILY MEN's attitudes toward leisure vary.

$\mathbb{K}^{\mathcal{T}}$	FAMILY MAN	SPORTS FAN	INTELLECTUAL	PRACTITIONER	leisure unimportant	leisure less important	leisure important	leisure rather important	leisure very important
1				×	×	×	×	×	×
2			×		×	×	×		
3		×			×				
4	×	×			×	×			
5			×	×	×	×	×	×	
6		×			×	×			
7		×			×				
8					×	×	×		
9			×	×	×	×	×	×	
10	×		×	×	×	×	×	×	×
11	×	×			×	×			
12		×			×				
13			×	×	×	×	×	×	
14			×	×	×	×	×	×	×
15		×			×	×	×		

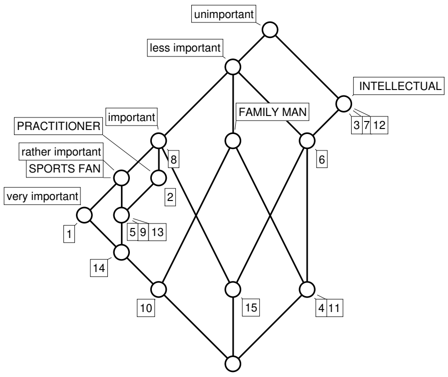

Fig. 6. Context *leisure activities* derived by the terminology *stereotypes* and corresponding concept lattice

At the same time, this concept lattice does not contain any information about a specific kind of sport. This is because they were considered non-relevant for the expert's, say sociological, theory in use. Thus, the attribute sports was not part of the chosen terminology.

4 Logical scaling based on relational contexts

The language SQL comprises several data types equipped with further structures. For example, for the construction of predicates like (importance ≤ 3) we utilized the fact that SQL offers the data type "integer" equipped with the natural order. Theoretically, it would be possible to define an equivalent predicate using only combinations of $\wedge$ and $\vee$. Nevertheless, we prefer the former since the order is constituent of our natural understanding of the notion "importance" and its values. That is why we want to have relations as elements of our language.

Introducing relations as part of the formalization and the language, we can express explicitly the implicit structures of value sets on which scaling should always be based. Therefore, we represent data in a *relational context* in which all implicitly given relations are formalized explicitly. A more detailed motivation of this notion following the lines of the theory of measurement can be found in [Pre96].

Following [Pri96] we can define a *relational context* formally as a tuple $(G, M, (W, \mathcal{R}), \mathcal{I})$, where (G, M, W, I) is a many-valued context and $\mathcal{R}$ a set of relations on W.

Let us suppose we had specified the information given in [AL83] about the structure of the value sets by relations of our many-valued context *leisure activities*. Then we could have used them to construct predicates and would not have been restricted to the relations provided by SQL.

For example, let us look at the tree-like hierarchy with its five levels which is described by the scale $\mathbb{S}_{\mathsf{sports}} := (W_s, M_s, I_s)$. We can describe it by a family $(S_n)_{n=1,\ldots,4}$ of binary relations on the set $W_s \subseteq W$ of sports: two kinds of sports $v, w \in W_s$ are in relation S_n to one another if they belong to the same extension of a concept of level n.

$$(v, w) \in S_n :\Longleftrightarrow \exists m \in M_s : \; v, w \in m^{I_s} \text{ and } |m^{I_s}| = n$$

These relations S_n can be interpreted as similarity of degree n, i. e. as "identical" (S_1), "very similar" (S_2), "rather similar" (S_3), and "similar" (S_4). Providing these relations, we are able to generate the predicate

$$(\mathsf{sports}\ S_3\ \mathrm{su})$$

which is satisfied by all persons who like a kind of sport which is "rather similar" to surfing, i. e. all persons who are surfing, skiing or sailing. Note that this predicate can be constructed without specifying sailing and skiing explicitly. In SQL, the predicate would be implemented by the following command:

`SELECT` Person `FROM` *activities* `WHERE` sports `IN`
(`SELECT` W_{s1} `FROM` S_3 `WHERE` $W_{s2} = \mathrm{su}$)

For the construction of predicates, we can also use n-ary relations for $n \geq 2$, in particular operations on values of different attributes. As an example, we imagine an extension of our context *leisure activities* with five many-valued attributes $D_1, D_2, ..., D_5$. They specify, for each person, the disbursements of the person's money for leisure activities in five succeeding weeks. Then, the expert may decide that the importance of leisure should not be measured by the subjective judgement that is given by the many-valued attribute importance, but by the average disbursement for leisure activities. Therefore, the expert could describe an attribute called "leisure important" by the predicate below which is satisfied by all persons whose average disbursement for leisure activities is under 15 \$ per week:

$$(\frac{1}{5}(D_1 + D_2 + D_3 + D_4 + D_5) \leq 15)$$

This example shows again that the choice of predicates depends on the specific issue of the data analysis. Such a description of "importance of leisure" is certainly determined by the interest of market researchers of finding potential clients with adequate purchasing power and will. Nevertheless, logical scaling always provides the possibility to discuss these decisions because they are listed in the terminology.

5 Discussion

To sum up, formal concept analysis provides different methods to transform many-valued contexts into one-valued contexts from which we can explore conceptual patterns. Both are useful for the treatment of real-world problems.

Conceptual scaling is a well established method which is frequently applied. Supported by the software tools TOSCANA and ANACONDA, the method allows a global view on the conceptual structure of the data context with regard to the inherent structures of the value sets and the research questions.

Logical scaling, on the other hand, is suitable for more specific issues. The expert decides which of the attributes and attribute values of the many-valued context are relevant to his analysis and then declares explicitly how the predicates are to be constructed. This method has several advantages. Firstly, we can construct quite complex predicates by using relations, disjunctions and other elements of our formal language. Secondly, specifying terminologies is more intuitive than defining conceptual scales for people who are not accustomed to work with concept lattices. That is why it might be a great convenience to combine both methods and to use logical scaling for the construction of concrete conceptual scales.

And finally, the explicit declaration of the terminology in view offers the possibility to discuss the decisions at all times because they are always visible. Thus, both scaling methods serve the purpose to support the communication about the data and their conceptual patterns instead of providing "results" automatically.

The choice of the formal language was determined by practical considerations to emphasize the applicability to real-world problems. SQL has a simple structure which is comprehensible without detailed explanations about the syntax and semantics. Additionally, SQL makes it possible to integrate a logical scaling tool into ANACONDA because the program is based on the relational database management system Microsoft Access which has SQL as its query language.

In [Pre96], where the idea of logical scaling was first discussed, we did not use SQL but an attribute logic that is influenced by description logics, a family of knowledge representation languages presented in [SS91]. For this language, questions of algorithmic treating are well investigated.

It might also be interesting to use conceptual graphs and the implicit hierarchy on conceptual graphs to structure terminologies as Sowa hints in [So96]. This may lead to an interesting cooperation between conceptual graphs and formal concept analysis.

References

[AL83] K. Ambrosi, W. Lauwerth: Ein Klassifikationsverfahren bei qualitativen Merkmalen, in: I. Dahlberg, M. R. Schader (ed.): Automatisierung in der Klassifikation, Indeks Verlag, Frankfurt 1983, 151 – 160

[Fo77] H. T. Forst: Anwendung der Cluster-Analyse zur Typisierung des Freizeitverhalten von Jugendlichen, in: H. Späth: Fallstudien Cluster-Analyse, Oldenbourg Verlag, München Wien 1977, 161 - 178

[GSW86] B. Ganter, J. Stahl, R. Wille: Conceptual measurement and many-valued contexts, in: W. Gaul, M. Schader: Classification as a tool of research, Elsevier Science Publishers, North-Holland 1986, 169 - 176

[GW89] B. Ganter, R. Wille: Conceptual scaling, in: F. Roberts (ed.): Applications of combinatorics and graph theory to the biological and social sciences, Springer-Verlag, New York 1989, 139 – 167

[GW96] B. Ganter, R. Wille: Formale Begriffsanalyse. Mathematische Grundlagen, Springer-Verlag, Berlin – Heidelberg 1996

[In91] Informix: Informix Guide to SQL Tutorial, Bohannon Drive 1991

[So92] J. F. Sowa: Conceptual Graphs Summary, in:T. E. Nagle et. al. (ed.): Conceptual Structures. Current research and practice, Proceedings, 2nd International Conference on Conceptual Structures, ICCS 1992, 1 – 51

[So96] J. F. Sowa: Processes and Participants, in: P. W. Eklund / G. Ellis / G. Mann (ed.): Conceptual Structures. Knowledge representation as Interlingua, Proceedings, 4th International Conference on Conceptual Structures, ICCS 96, Sydney, Springer, Berlin New York 1996, 1 – 22

[Pre96] S. Prediger: Symbolische Datenanalyse und ihre begriffsanalytische Einordnung, Staatsexamenarbeit, FB Mathematik, TH Darmstadt 1996

[Pri96] U. Priß: The formalization of WordNet by methods of relational concept analysis, in: C. Fellbaum (ed.): WordNet – An electronic lexical database and some of its applications, MIT-Press 1996

[SS91] M. Schmidt-Schauß, G. Smolka: Attributive concept descriptions with complements, in: Artificial Intelligence 48 (1991), 1 – 26

[Wi82] R. Wille: Restructuring lattice theory: an appraoch based on hierarchies of concepts, in: I. Rival (ed.): Ordered sets. Reidel, Dordrecht–Boston 1982, 445 – 470

[Wi92] R. Wille: Concept Lattices and Conceptual Knowledge Systems, in: Computers & Math. Applications, vol. 23, no. 5 – 9, 1992

This article was processed using the LaTeX macro package with LLNCS style